Calculations of Spectroscopical Properties of Extended Systems

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch–naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Samuele Giani

von

Airolo TI

Promotionskomitee

Prof. Dr. Jürg Hutter (Vorsitz)

Prof. Dr. Peter Hamm

Prof. Dr. Stefan Seeger

Prof. Dr. Paolo Carloni

Zürich, 2014

Everybody makes mistakes.
That's why they put erasers on pencils

(Carl Carlson)

Should one look at right and wrong as ethical questions? That is the problem. Marcuse says Hegel's "Philosophy of Right" does not assign a moral category to "wrong". Free will inevitably causes wrong.

You have to be prepared to reconsider right and wrong. Because basically those are just terms that express a horrible struggle, part of an equation of pure dialectic. Strip the sentiment out and try to come to terms with historical forces external to us and indifferent to moral category.

character of Yvonne

taken from the movie
"Munich", Steven Spelberg (2005)

Abstract

Nowadays massively parallel computers are applied to the study of a multitude of biomolecular phenomena. In this respect, the work presented here is relying on computer simulations and linked to two main goals: to advance the understanding of properties related to X–rays and NMR spectroscopy, and to explore novel ideas in massively parallel machine architecture and software.

Increased computational power translates into an increased ability to validate the models used in simulations and, in the context of life sciences, to probe biological processes at the microscopic level over the appropriate time periods. A crucial component of this research is the connection of the simulation outcome to the experimental data available.

This thesis is divided in three parts. At first, several basic arguments are introduced: aspects on which subsequent theoretical developments are based upon. These latter are indeed the subject of the two distinct remaining parts. The first is related to a phenomenon occurring by energetic promotion of core electrons (NEXAFS), while the second is inherent to excitation of the nuclear spin (NMR).

Both of these spectroscopies involve calculations of properties in which, by nature, electronic density around the nucleus plays an important role.

The explicit representation in the calculation of the core electrons, also known as inner shell electrons, is appropriately and efficiently described within the GAPW framework. This extension is very important to reproduce with high accuracy NEXAFS and NMR properties, because they depend critically on the details of the wavefunctions in the core regions.

Accurate calculations of properties of materials that can be directly compared with experimental results are of great importance, since they supply an essential help in the interpretation of measured data. Moreover, structure–property correlations that are never resolved in experiments on a molecular level, could be quantified and unknown features of scientific and technological interest can be predicted.

Zusammenfassung

In der heutigen Zeit werden massiv–parallel Computer für Studien einer Vielzahl von biomolekularen Phänomenen angewandt. In dieser Hinsicht basiert die hier vorgestellte Arbeit auf Computersimulationen und ist an zwei Hauptziele geknüpft: die Erweiterung des Verständnis in Bezug auf die Eigenschaften von Röntgenstrahlen und der NMR Spektroskopie sowie die Erforschung neuer Konzepte im Bereich der Architektur und Software von massiv–parallel Rechnern.

Erhöhte Computerleistung führt zu einer erhöhten Fähigkeit, die in Simulationen angewandten Modelle zu validieren und im Zusammenhang von Naturwissenschaften biologische Prozesse auf der mikroskopischen Ebene über die angemessenen Zeiträume zu untersuchen. Eine wesentliche Komponente dieser Forschung ist die Verbindung des Simulationsergebnisses zu den verfügbaren Experimentaldaten.

Die vorliegende Arbeit ist in drei Teile gegliedert. Zunächst werden mehrere grundlegende Argumente vorgestellt: Aspekte, auf denen nachfolgende theoretische Entwicklungen basieren. Und diese sind Gegenstand der zwei verbleibenden und unterschiedlichen Teile. Der eine bezieht sich auf ein Phänomen, welches bei energetischer Förderung von Kernelektronen (NEXAFS) auftritt, während der andere Teil sich auf die Anregung von Kernspin (NMR) bezieht.

Beide dieser Spektroskopien beinhalten Kalkulationen von Eigenschaften, in denen die Elektronendichte um den Nukleus naturgemäss eine wichtige Rolle spielt.

Die explizite Darstellung der Berechnung der Kernelektronen, auch als Innerschalenelektronen bekannt, wird im Rahmen der GAPW Methode an-

gemessen und effizient beschrieben. Es ist sehr wichtig, diese Erweiterung mit hoch präzisen NEXAFS und NMR Eigenschaften zu reproduzieren, da sie stark von den Details der Wellenfunktionen in den Kernregionen abhängig sind.

Genaue Berechnungen von Materialeigenschaften, die unmittelbar mit Experimentalergebnissen verglichen werden können, sind von grosser Wichtigkeit, da sie in der Interpretation von Messdaten eine wesentliche Hilfe darstellen. Darüber hinaus könnten Struktur-Eigenschaft Zusammenhänge, die in Versuchen auf Molekularebene nicht geklärt werden können, quantifiziert und unbekannte Merkmale von wissenschaftlichem und technologischem Interesse prognostiziert werden.

Table of Contents

Abstract **i**

I Basic Theory **1**

Introduction **3**

1 Background Topics **5**

1.1 *Ab Initio* Philosophy 5

1.2 Sampling From Ensembles 8

References 10

2 GPW and GAPW Formalisms **11**

2.1 The Method 11

References 19

II X–Ray Absorption Spectroscopy **21**

3 Methodology of NEXAFS calculations **27**

3.1 Slater Transition Potential Method 27

3.2 Spectra determination procedure 30

References 33

4 C, N, O and S NEXAFS Calculations of Solvated Peptides 35

4.1 Introduction 35

4.2 Calculations 36

4.3 Glycine 38

4.4 Peptides 50

4.5 Conclusions 62

References 65

III Nuclear Magnetic Resonance 69

5 Magnetic Linear Response Properties Calculations With the GAPW Method 71

5.1 Theory 73

5.2 Calculation of the Induced Current Densities 74

5.3 Calculation of the Chemical Shift Tensor 77

5.4 Results and Discussion 78

5.5 Summary 85

References 87

6 Aqueous Simulations of NMR Spectra of Amino Acids and Peptides 93

6.1 Introduction 93

6.2 Computational Details 96

6.3 Simulation of Amino Acids Spectra 99

6.4 Solvated Polypeptide: β–Hairpin BIV2 121

6.5 Conclusions 130

References 132

Acknowledgments 139

Curriculum Vitæ 141

Part I

Basic Theory

Introduction

Present–day computer technology has changed profoundly the way in which scientific research is conducted. Computers have speeded up considerably the growth of theoretical and experimental research, since they perform in an automatic way and at an incredible speed calculations that would otherwise have taken far longer to be executed. However, their influence does not stop here and computers have become real research instruments on which scientific investigations are conducted. This is due mainly to the maturity of computer simulations, that is to say the ability to track on the computer the properties of model systems describing with ever greater accuracy the behavior of complex systems. This has led to the birth of a new way of doing science that is between theory and experiment, namely computer simulation. As in purely theoretical research, one works with a model, but the approach to problem solution is similar to the experimental one. This is also reflected in the language used: one performs computer experiments, measures quantities, worries about signal to noise ratio, etc., but the experiments or simulations are done on the computer rather than in the laboratory.

First principles calculations are widely used as the essential tool to study the properties of matter at atomic scale and to support experimental research in the interpretation of measured data and phenomenological evidences. Even more useful is the combination of properties calculation with molecular dynamics (MD), for the investigation of different thermodynamic conditions and reaction mechanisms. A multitude of *ab initio* methods are available to calculate those properties at different level of accuracy. However, due to the prohibitive computational costs of some approaches, calculations are sometimes restricted to small samples, containing few non–hydrogen atoms and where long range environmental effects cannot be taken into account. Because of the demanding computational costs of such simulations, methods

based on Density Functional Theory (DFT), in its self–consistent formulation, have become more and more popular, in particular in combination with the plane waves (PW) basis set and the pseudopotential (PP) approximation. This approach has proved to be efficient and accurate for the calculation of total energies and electronic properties like bond lengths and angles, as well as vibrational frequencies, *i.e.* those quantities that depend on the electronic structure outside the core. But what happens if the property of interest is sensitive to the electronic structure in the core? For a proper comparison with experiments, it is then crucial to appropriately model the core electrons, and to extend the scope of simulations to treat large systems and long time scales. For this reason, our aims are devoted towards the development of DFT–based linear scaling approaches, which may open the way to quantum chemistry prediction of macromolecules and extended condensed phase systems.

CP2K

The CP2K code is a multipurpose atomistic code, especially devoted to perform DFT electronic structure calculations. In principle conceived to tackle the bottlenecks that hindered efficient large scale *ab initio* molecular dynamics. It has also been extended to deal with Monte Carlo sampling and classical force fields molecular mechanics. It benefits efficiently from parallel computer architectures and it is distributed under the GNU General Public License (GPL), making it an open source software.

Chapter 1

Background Topics

1.1 *Ab Initio* Philosophy

By mentioning large scale calculations, let us express what this definition is referring to. Explicitly solvated systems, biomolecules like proteins or DNA, interfaces (solid/solid, solid/liquid, liquid/vapor) or defective solid state systems (dislocations or impurities) are current target applications for electronic structure calculations. Phenomena one might study are enzymatic catalysis, electron transfer processes, molecular electronics or properties in general. If those are prototypical examples, a fundamental question still remains: what means this in terms of number of atoms? Is this 100, 1 000 or 10 000? Modeling these type of systems with acceptable accuracy might require the use of large samples (many atoms in the simulation box) in order to avoid artifacts induced by size effects. The optimal size for a model is determined by the specific system and the addressed properties.

Moreover, investigating dynamical evolution, transformations and kinetics of the system require proper sampling with techniques like molecular dynamics or Monte Carlo. This can be achieved only by generating a large number of configurations, *i.e.* extending the sampling over long trajectories.

At this point the time domain is important, emphasizing that doing one electronic structure calculation is not enough, and the correct strategy is to perform statistical sampling. Large scale thus also means lot of calculations.

The simulation of the electronic structure of such systems, is often frustrated by the fast increase of computational costs with the system size N. For traditional algorithms, indeed, the scaling behavior ranges from $\mathcal{O}(N^8)$, for highly accurate couple cluster theory, to $\mathcal{O}(N^3)$, for independent particle models like Density Functional Theory (DFT). The Kohn–Sham (KS) formulation of DFT maps the system of interacting particles to an ensemble of electrons that only interact through the total electron density $\rho(\mathbf{r})$. Correlation effects are implicitly included via the exchange–correlation (XC) functional, and the ground state energy is obtained by the self consistent minimization of the total energy functional E with respect to a set of orthogonal, one electron wavefunctions $\varphi_i(\mathbf{r})$. DFT provides a typical accuracy of post–Hartree–Fock methods (e.g. MP2) at a comparable cost of an HF calculation. In addition, the possibility to calculate atomic forces in a straightforward way makes DFT the method of choice for *ab initio* MD simulations.

1.1.1 Density Functional Theory and Kohn–Sham Ansatz

The fundamental of DFT is that any property of a system of many interacting particles can be viewed as a functional of the ground state density $\rho(\mathbf{r})$ that determines all the information in the many–body wavefunctions for the ground and all excited states. The existence proofs for such functionals, given in the original works of Hohenberg and Kohn and of Mermin, are disarmingly simple [3]. However, they provide no guidance whatsoever for constructing those functionals, and no exact functionals are known for any system of more than one electron. DFT would have been a marginal curiosity if it were not for the ansatz made by Kohn and Sham (KS), which provided a way to make useful approximate ground state functionals for real systems of many electrons [5].

The starting point of the following discussion is non–relativistic quantum mechanics as formalized via the time–dependent Schrödinger equation. The theoretical approach we will rely on is, as mentioned, Kohn–Sham DFT [4] and hereafter are highlighted just the basics.

A molecular orbital is expanded in a set of basis set (φ_α) functions, as

$$\Phi_i(\mathbf{r}) = \sum_\alpha c_{\alpha i}\, \varphi_\alpha(\mathbf{r}) \tag{1.1}$$

the $c_{\alpha i}$ are the variational parameter to be optimized.

Where the basis is not orthogonal, the overlap matrix reads

$$\mathbf{S}_{\alpha\beta} = \int \varphi^\star_\alpha(\mathbf{r})\, \varphi_\beta(\mathbf{r})\, d\mathbf{r}\ . \tag{1.2}$$

The orbitals have to fulfill the orthogonality condition, this is written within the overlap matrix in the form

$$\int \Phi^\star_i(\mathbf{r})\, \Phi_j(\mathbf{r})\, d\mathbf{r} = \sum_{\alpha\beta} c^\star_{\alpha i}\, \mathbf{S}_{\alpha\beta}\, c_{\beta j} = \delta_{ij}\ . \tag{1.3}$$

This is a dramatic simplification since the minimization with respect to all possible many–body wavefunctions Φ is replaced by a minimization with respect to a set of orthonormal one–particle functions, this latter being seen as an auxiliary independent–particle system that can be solved more easily. From the orbitals coefficients the density matrix $\mathbf{P}$ can be calculated

$$\mathbf{P}_{\alpha\beta} = \sum_i f_i\, c^\star_{\alpha i}\, c_{\beta i}\ , \tag{1.4}$$

where f_i are integer occupation numbers and, via the density matrix, the electron density $\rho(\mathbf{r})$ can be obtained from a single Slater determinant built from the occupied orbitals

$$\rho(\mathbf{r}) = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\, \varphi^\star_\alpha(\mathbf{r})\varphi_\beta(\mathbf{r})\ . \tag{1.5}$$

The KS recipe tells us what to do to calculate the ground–state energy, namely to minimize over all possible realization of the c_i the following energy expression, where there is the kinetic energy, that is directly a functional of the orbitals, the external energy as interaction of the electrons with the nuclei, the Hartree energy in the form of the classical Coulomb energy, and

the exchange–correlation energy. This minimization has to be carried out subjected to the orthogonality constraint (Eq. (1.3)).

$$E = \mathrm{Min}_{\,c_i} \left[E_{\mathrm{kin}}(c_i) + E_{\mathrm{ext}}(\rho) + E_{\mathrm{H}}(\rho) + E_{\mathrm{xc}}(\rho) \right] \tag{1.6}$$

This can be performed with different methods, the standard one would be a fix point iteration with diagonalization, that means by taking energy expression with the constraint, do a variational minimization and what one gets is an eigenvalues equation, the KS equations, which are one–electron equations involving an effective one–particle Hamiltonian with local potential

$$\sum_{\beta} \mathbf{H}_{\alpha\beta}\,(c^n)\; c_{\beta i}^{n+1} = \varepsilon_i\, c_{\alpha i}^{n+1} \tag{1.7}$$

after plugging in trial c^n, the resulting eigenvectors will be the next orbitals. Doing this iteratively until self consistency one gets at the end the KS energy. Another option would be a direct optimization with constraint, calculating the gradient with the constraint and directing optimize for the minimum. What has to be calculated in both cases is the KS matrix

$$\mathbf{H}_{\alpha\beta}(c) = \langle\, \varphi_\alpha \mid \widehat{T} + \widehat{V}_{\mathrm{ext}} + \widehat{V}_{\mathrm{H}}(n) + \widehat{V}_{\mathrm{xc}}(n) \mid \varphi_\beta \,\rangle\,. \tag{1.8}$$

in this equation the matrix element $\mathbf{H}_{\alpha\beta}(c)$ is calculated for the operators between two basis functions and the operators are respectively the kinetic energy, the external operator, the coulomb potential and the exchange–correlation potential, in analogy to Eq. (1.6). This Hamiltonian nevertheless embodies the electronic many–body effects by virtue of the XC potential.

1.2 Sampling From Ensembles

The goal pursued from molecular dynamics is to model the detailed microscopic dynamical behavior of many different types of systems as found in chemistry, physics or biology [2]. Physical properties of molecular or atomic systems are measured or estimated by statistically meaningful ensemble averages under specific thermodynamic conditions.

Nuclei are treated as classical particles subjected to the laws of classical mechanics. Quantum effects need to be taken into account only for special cases of light and hot particles. This is a very good approximation for molecular systems as long as the properties studied are not related to the motion of light atoms (*i.e.* hydrogen) or vibrations with a frequency ν such that $\nu > k_B T$.

Let's consider a system of N particles moving under the influence of a potential function. The forces on the particle are derived from the potential and the equations of motion evolve according to Newton's second law. The Hamilton's equation of motion are time reversible and the total energy is a constant of motion. Properties are important to establish a link between molecular dynamics and statistical mechanics, the latter connecting the microscopic details of a system and the physical observables such as equilibrium thermodynamic properties. Statistical mechanics is based on the Gibbs ensemble concept: that is, many individual microscopic configurations of a very large system lead to the same macroscopic properties, implying that it is not necessary to know the precise detailed motion of every particle in a system in order to predict its properties. It is sufficient to simply average over a large number of identical systems, each in a different configuration; *i.e.* the macroscopic observables of a system are formulated in term of ensemble averages [1].

Statistical ensembles are characterized by fixed values of thermodynamic variables, and regarded as experimental control parameters that specify the conditions under which an experiment is performed. Therefore, a dynamical trajectory (the position and momenta of all particles over time) will generate a series of states corresponding to a given ensemble, lying on a defined hypersurface in the phase space. The assumption that a system, given an infinite amount of time, will explore the entire hypersurface is known as the ergodic hypothesis. Thus, under this hypothesis, averages over a trajectory of a system obeying Hamilton's equation are equivalent to averages over the given ensemble [2].

References

[1] M.P. Allen and D.J. Tildesley. *Computer simulations of liquids.* Claredon Press, 1987.

[2] D. Frenkel and Smit B. *Understanding molecular simulation.* Academic Press, second edition, 2002.

[3] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864, 1964.

[4] W. Koch and M.C. Holthausen. *A chemist's guide to density functional theory.* Wiley–VCH, 2000.

[5] W. Kohn and L.J. Sham. Self–consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133, 1965.

Chapter 2

GPW and GAPW Formalisms

Going towards large scale systems, linear scaling (LS) procedures are needed to reduce the computational work performed and make simulations feasible. In the context of mere electronic structure calculations, one has to tackle two distinct parts: calculate the KS matrix and optimize the molecular orbitals (MOS) by updating the electronic density within the self consistence cycle (see Eq. (1.7)). If both of this parts are calculated LS, the overall method is coherently $\mathcal{O}(N)$).

For the KS matrix this condition might be achieved by relying on a localized basis sets for the expansion of the MOS. If the basis set has finite support, it results localized and the KS setup should in principle be possible to be achieved in the LS regime.

2.1 The Method

A localized basis set is what is needed, and one choice are atomic orbital based BS, which mimic the atomic orbitals and they move with the atom when the latter is displaced. Typical basis sets of this kind are Gaussians (Gauss functions or GTO) or exponential functions (so called Slater functions). The electronic structure calculations performed within the QS module of `CP2K` relies on the expansion of the MOS in therms of GTO. The main advantage of this choice is that these functions can be efficiently integrated analytically over many quantum mechanical operators relying on

recurrence relations [11], and having a finite support the resulting KS and overlap matrix may have a rather high degree of sparsity.

As long as one have local operators, and it turns out to be the case for the kinetic energy operator and the potential in standard KS DFT, then the Hamiltonian matrix elements can be written like

$$\mathbf{H}_{\alpha\beta} = \int \widehat{H}(\mathbf{r})\ \varphi_\alpha(\mathbf{r})\varphi_\beta(\mathbf{r})\ d\mathbf{r} \tag{2.1}$$

the product $\varphi_\alpha(\mathbf{r})\varphi_\beta(\mathbf{r})$ between two basis functions is nonzero only for overlapping functions, and only those pairs have to be integrated. For functions with finite support then this product differs from zero only if φ_β and φ_α are centered on sufficiently close atoms. $\mathbf{H}_{\alpha\beta}$ results thus sparse [3].

In other words, if one enlarges the system, the number of basis functions accordingly increases, however the number of neighbors of a given atom is the same: Therefore, integrals of the type of Eq. (2.1) that need to be computed increases linearly with system size, and the sparsity of the KS matrix is also more evident. Hence, employing order one algorithms to calculate the KS matrix elements, obtaining the KS matrix results to be a $\mathcal{O}(N)$ procedure.

If the operators are of $\mathcal{O}(N)$, like the overlap matrix and the kinetic energy, this condition is satisfied. The only thing that has to be done is to calculate efficiently is the neighbor list, and how to be performed it is known from molecular dynamics algorithms [1]. The XC potential, if LDA or GGA DFT is used, is easy, but the electrostatic, so the potential coming from the charge distribution, is the difficult case. The term depending on the charge density, namely V_H, is not local, so the contributions to be computed extend beyond a short list of neighbors, and special treatment is required to keep the calculation efficient (*i.e.* LS).

To this purpose we use schemes based on the expansion of the charge density in plane waves (PW), in order to solve the Poisson equation in the reciprocal space. The standard approach, GPW, employs the PW as auxiliary basis set, the full charge density is collocated on a regular grid in real space and then transformed by FFT into the PW expansion.

In the GPW procedure all the contribution to the density needs to be accurately described on the real space grid, this latter has to be sufficiently

dense. Where the charge density includes rather hard terms (e.g. large exponents in the GTO representation) a proper PW expansion requires large energy cutoff. This results in an important loss in computational speed, in particular when the system includes hard elements like core states.

As an alternative, using the GAPW scheme [10], only the softer contribution of the density are collocated on the grid and then transformed into the PW expansion in reciprocal space. The remaining hard contributions, which are also the most localized terms close to the atomic centers, are collocated on local spherical grids, centered on each atom. The corresponding exchange–correlation and Hartree contributions to the KS matrix, are calculated by integration on these grids.

The Coulomb energy is given by the double integral, or integral over six dimensions, including the electronic density

$$E_{\mathrm{H}} = \frac{1}{2} \int\int \frac{\rho(\mathbf{r})\, \rho(\mathbf{r}')}{|\, \mathbf{r} - \mathbf{r}' \,|}\, d\mathbf{r}\, d\mathbf{r}' \tag{2.2}$$

The strategy adopted is to take the Fourier transform of Eq. (2.2), thus writing the total Coulomb energy in the following form, relying on a plane wave expansion (PW) of the total charge density [5]

$$E_{\mathrm{H}} = \frac{2\pi}{\Omega} \sum_{\mathbf{G}} \frac{\rho^*(\mathbf{G})\, \rho(\mathbf{G})}{\mathbf{G}^2} \tag{2.3}$$

The question is how to calculate the plane wave expansion of the total charge density when the wavefunctions are given in Gaussian functions in a way that is LS. In other words, we aim using for the density a different basis set that for the wavefunctions. For the φ having Gaussian functions and for ρ using plane waves. The important part is how to switch from a orbital in Gaussian representation to a density in plane wave representation. Once we have that, calculating the potential is easy, we would just to divide by $\mathbf{G}^2$.

The important point is that by introducing PW basis, we introduce a regular grid in real space and another in $\mathbf{G}$–space, and one can go to one to the other without loss of accuracy by using fast Fourier transforms (FFT) [5]. Going from one space to the other is almost LS, is $\mathcal{O}(n \log n)$, n being the number of grid points.

The algorithm for the Coulomb potential is schematically depicted:

$$\mathbf{P} \rightarrow \rho(\mathbf{R}) \underbrace{\xrightarrow{\text{FFT}} \rho(\mathbf{G}) \rightarrow V_{\text{H}}(\mathbf{G}) = \frac{\rho(\mathbf{G})}{\mathbf{G}^2} \xrightarrow{\text{FFT}}}_{\mathcal{O}(n \log n)} V_{\text{H}}(\mathbf{R}) \rightarrow \boldsymbol{V} \tag{2.4}$$

The procedure is performed in this manner: starting by having a density matrix obtained from the coefficients c_i, the density is then collocated on the real space grid, and then transformed by FFT into the representation in $\mathbf{G}$ space. The potential is then calculated by dividing the density by $\mathbf{G}^2$, and backtransformed via FFT onto the real space grid. Numerical space integrations gives the corresponding KS matrix elements [9].

The FFT part is $\mathcal{O}(n \log n)$ scaling as already pointed out. What's left is to show that we can do the remaining part in LS. $\rho(\mathbf{r})$ is collocated on the real space grid: the density is written as a sum over $\mathbf{P}$ times the product of basis functions. This has to be made for each point on the grid. These are local functions, so they are nonzero only at special points. The product is only nonzero when the two functions overlap, having the same properties mentioned before, and only if this function have an overlap the calculation is done, and only in the part where they overlap. It is just important that the screening for the sums is carried out effectively.

The XC energy is also calculated via a numerical integration on the real space grid. The XC potential is calculated on the grid and and then added to the V_H potential. Afterward the numerical integrations is performed only once.

The general drawback of this method is the same one in common to all PW methods. The density close to the atoms can vary rather quickly, meaning that a high cutoff for the BS expansion is needed. The standard way out is to rely on pseudopotentials (PP): typically only valence electrons are treated explicitely while core electrons are frozen and the interaction between valence and core is mediated by the PP. Hence, only the smoother density of the valence electrons need to be described in the PW expansion, thus limiting the size of the PW basis set to functions with kinetic energy below a few hundred of Ry (cutoff 300 – 500 Ry). In GPW the PP used are of the dual

space type form [4]. The use of the PP is a limitation we would like to overcome, because we aim to perform all electrons (AE) calculations. There are cases where also core electrons need to be explicitely described. In order to avoid the expansion in PW of the hard core density, dual basis set method have been devised.

There is a similar method, introduced by Blöchl, under the name projector augmentation method (PAW) [2], that was brought into practice to be able to do calculations without PP, from which the GAPW is inspired and some concepts are borrowed.

GAPW

The idea is that the density can be separated in contributions with different character, that can be treated in different manners. The important aspect to underline is that all the computational grids and the respective representations of the different contributions to the density are defined in all space, avoiding to be separated in exclusive parts [10, 7]. The space is partitioned: a part close to the atoms (A) and another far away from atoms. The partition in outer space is assumed to have a low cutoff and can be expanded in PW: this is called interstitial region (I). Thus for I the density is written as a density that is assumed to be smooth, hence the expansion in PW. In A we expand the density on a radial grid. To do this in a conventional way we would have in principle to match the densities at the borders. What is wanted is to avoid matching procedures and what is done is that this atomic region is added also to the interstitial region: the density of the A region might go outside the A region, so we have to subtract another density that just compensates this. The same inversely: for the I region, $\tilde{\rho}(\mathbf{r})$ is allowed to extend within the A region but we require that this soft atomic density contribution has to be counterbalanced by another therm. By adding the parts, what is gotten is one expansion of the density valid in all space: so no more boundaries. This is a function that everywhere in space is truly representing $\rho(\mathbf{r})$. Now we have the full density separated in a smooth density that is expanded in PW and atomic densities that are expanded in local functions (Gaussians). For a more detailed and comprehensive description, the reader is addressed to the review [12].

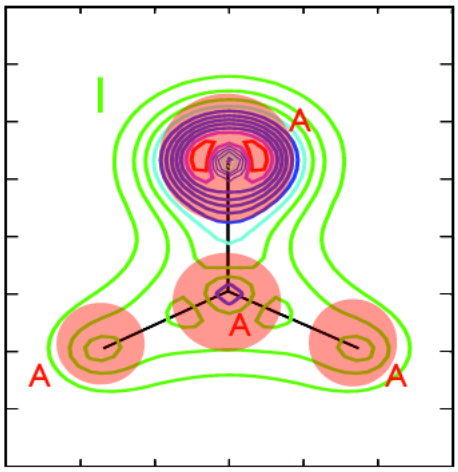


Figure 2.1: Electronic representation: contour plot of the ground–state density of formaldehyde ($H_2C{=}O$), in the plane of the molecule.

Conditions on density representations:

▶ Interstitial Region (I)

$$\rho(\mathbf{r}) = \tilde{\rho}(\mathbf{r}) + \sum_A \rho_A(\mathbf{r}) - \sum_A \tilde{\rho}_A(\mathbf{r}) \tag{2.5}$$

$$\rho_A(\mathbf{r}) = \tilde{\rho}_A(\mathbf{r}) \tag{2.6}$$

▶ Atomic Regions (A)

$$\rho(\mathbf{r}) = \rho_A(\mathbf{r}) + \tilde{\rho}(\mathbf{r}) - \tilde{\rho}_A(\mathbf{r}) \tag{2.7}$$

$$\tilde{\rho}(\mathbf{r}) = \tilde{\rho}_A(\mathbf{r}) \tag{2.8}$$

▶ Everywhere in Space

$$\rho(\mathbf{r}) = \tilde{\rho}(\mathbf{r}) + \sum_A \rho_A(\mathbf{r}) - \sum_A \tilde{\rho}_A(\mathbf{r}) \tag{2.9}$$

The basis set used is composed by primitives that are monomials and exponential functions,

$$g_m(\mathbf{r}) = x^{m_x} y^{m_y} z^{m_z} e^{-\alpha_m \mathbf{r}^2} \tag{2.10}$$

The d's in Eq. (2.11) are not allowed to vary and our KS orbitals are linear combinations of this Gaussian functions. To get a smooth basis, all the primitives that have an α_m exponent higher than a cutoff threshold are just dropped. Having a function that is a superposition of Gaussians, after elimination of all the hard Gaussians, what's left is a soft function. We use this function to expand the density and this will be $\tilde{\rho}$.

$$\varphi_\alpha(\mathbf{r}) = \sum_m d_{m\alpha}\, g_m(\mathbf{r}) \tag{2.11}$$

$$\downarrow \quad \text{drop all } g_m \text{ with } \alpha_m > \alpha_{max}$$

$$\tilde{\varphi}_\alpha(\mathbf{r}) = \sum_{m'} d_{m'\alpha}\, g_{m'}(\mathbf{r}) \tag{2.12}$$

Thus the smooth density is constructed from the full density by omitting primitives with large exponents in Eq. (2.10).

For the atomic sites, we use the same primitives but this time not contracted to expand the full atomic density, and if we drop again the same exponents, we get the smooth density. This ensures the conditions pointed out before, namely that $\tilde{\rho}_A$ is equal ρ_A outside certain radius by construction. The same is true for the other density.

Atom centered densities ρ_A and $\tilde{\rho}_A$ are constructed by projecting all the contracted functions φ_a onto Gaussians localized on the atom A.

At this point the expansion of the density has to be plugged in the energy expression.

Hartree and XC terms are evaluated for the overall soft and for each atomic density.

The outlined principle of summing the contributions can be extended to the calculation of XC energy, this latter being just the sum of the three terms. This allows to calculate each part of the energy independently and there are no cross terms between two types of densities.

$$E_{\mathrm{xc}}(\rho) = E_{\mathrm{xc}}(\tilde{\rho}) + \sum_A E_{\mathrm{xc}}(\rho_A) - \sum_A E_{\mathrm{xc}}(\tilde{\rho}_A) \tag{2.13}$$

For the Hartree energy this is rather complicated [10], and we will just outline the result.

$$E_{\mathrm{H}}(\rho) = E_{\mathrm{H}}(\tilde{\rho}) + \sum_A E_{\mathrm{H}}(\rho_A) - \sum_A E_{\mathrm{H}}(\tilde{\rho}_A) + \text{multipole terms} \tag{2.14}$$

In fact, Blöchl derivations show that one have to calculate the Coulomb energy of the smooth density, add the Coulomb energy of the atomic densities non interacting each only with itself, subtracting the Coulomb energy of the smooth atomic densities. Multipole terms accounts for this correction, these latter originate because of the non–local nature of the operator giving rise to connections between the two parts.

Validation

A large number of calculations have been performed with both the GPW and GAPW scheme, and by a direct comparison to other quantum chemistry codes the quality of the results has been assessed [6, 13]. Validations show that total energies and geometrical parameters (bond lengths and bond angles) are reproduced with high precision. Also satisfactory are the values of harmonic frequencies. In summary, the accuracy of the method is noteworthy.

From a computational point of view, this approach allows high efficiency calculations with an early onset of linear scaling. One for all auxiliary basis, the possibility to use large Gaussians basis sets, in particular with a high angular momentum. Nuclear gradients are also efficiently calculated, unraveling the possibility to perform *ab-initio* molecular dynamics at an affordable computing cost [8].

References

[1] M.P. Allen and D.J. Tildesley. *Computer simulations of liquids*. Claredon Press, 1987.

[2] P. Blöchl. Projector augmented–wave method. *Phys. Rev. B*, 50:17953, 1994.

[3] S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085, 1999.

[4] S. Goedecker, M. Teter, and J. Hutter. Separable dual–space Gaussian pseudopotentials. *Phys. Rev. B*, 54:1703, 1996.

[5] J. Hutter and D. Marx. Proceeding of the february conference in Jülich. In J. Grotendorst, editor, *Modern methods and algorithms of quantum chemistry*, Jülich, 2000. John von Neumann Institute for Computing.

[6] M. Iannuzzi, T. Chassaing, T. Wallman, and J. Hutter. Ground and excited state density functional calculations with Gaussian and augmented plane waves method. *Chimia*, 59:499, 2005.

[7] M. Krack and M. Parrinello. All–electron *ab–initio* molecular dynamics. *Phys. Chem. Chem. Phys.*, 2:2105, 2000.

[8] I.-F. W. Kuo, C. J. Mundy, M. J. McGrath, J. I. Siepmann, J. VandeVondele, M. Sprik, J. Hutter, B. Chen, M. L. Klein, F. Mohamed, M. Krack, and M. Parrinello. Liquid water from first principles: Investigation of different sampling approaches. *J. Phys. Chem. B*, 108(34):12990–12998, 2004.

[9] G. Lippert, J. Hutter, and M. Parrinello. A hybrid Gaussian and plane wave density functional scheme. *Mol. Phys.*, 92:477, 1997.

[10] G Lippert, J. Hutter, and M. Parrinello. The Gaussian and augmented–plane–wave density functional method for *ab initio* molecular dynamics simulations. *Theor. Chem. Acc.*, 103:124, 1999.

[11] S. Obara and A. Saika. Efficient recursive computation of molecular integrals over cartesian gaussian functions. *J. Chem. Phys.*, 84(7):3963–3974, 1986.

[12] J. VandeVondele, M. Iannuzzi, and J. Hutter. Large scale condensed matter calculations using the gaussian and augmented plane waves method. In K. Binder, G. Ciccotti, and M. Ferrario, editors, *Computer simulations in condensed matter systems. Volume 1: fundamental techniques & approaches (lecture notes in physics, Vol. 703)*, page 287, 2005.

[13] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter. Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comp. Phys. Comm.*, 167:103, 2005.

Part II

X–Ray Absorption Spectroscopy

Physical Process

X–ray Diffraction

Since the beginning of the 20^{th} century, X–rays have been used to determine not only the equilibrium structure of crystalline materials, but also the electron density of molecules. In a crystal, atoms are arranged in periodic lattices. The incident X–ray experiences a three–dimensional structural arrangement, in the direction of constructive interference, a well–defined pattern of scattered beam intensities is observed, which satisfies Braggs law for diffraction. This technique offers the following advantages: a direct connection between atomic positions and the scattering amplitude can be obtained via the Fourier relation, and the visualization of electron densities, including the global structure, is possible.

X–ray Absorption Spectroscopy (XAS)

XAS measures the absorption of X–rays as a function of incident photon energy E. In a common spectra representing the absorption as a function of the photon energy the following general features can be observed (as sketched in Fig. (2.2)): an overall decrease in the absorption with increasing energy and the presence of saw–toot–like structures with a sharp rise at discrete energies, called absorption edges. The energy position of these transitions are unique to a given absorbing atom. They occur near the ionization energy of inner shell electrons and contain spectral signals due to core–to–unoccupied valence orbital transitions and core–to–continuum transitions [7, 3].

At energies above the edges, an oscillatory structure appears for molecules and solids that modulate the otherwise smooth absorption profile, typically

by a few percent of the absorption edge jump. These features, which are absent for single atoms in the gas phase, contain precise structural information related to interatomic distances and coordination numbers. XAS is divided into two domains: near–edge X–ray absorption fine structure (NEXAFS) for bound states and low–energy resonances [10, 4], and extended X–ray absorption fine structure (EXAFS) when the outgoing electron is well above the ionization continuum [8].

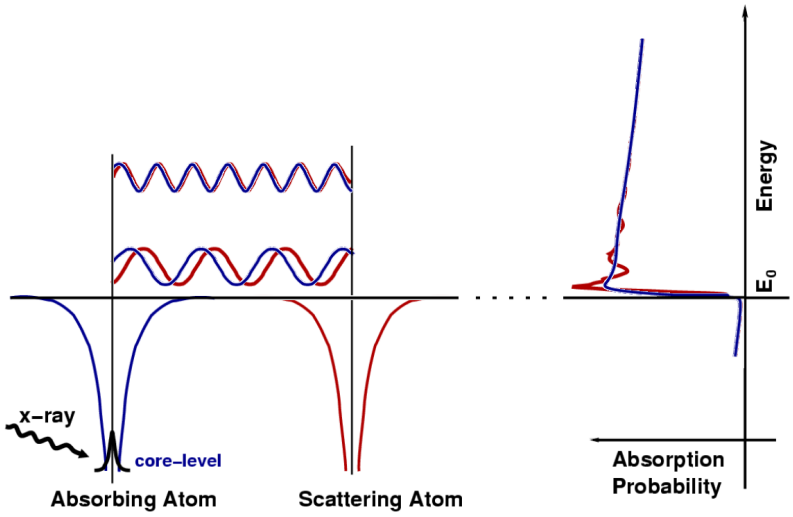


Figure 2.2: Cartoon of X–ray absorption process: Schematic potentials (left) and the absorption spectra (right). Because of the photoelectric effect, when a photon has the energy of a tightly bound core electron level, E_0, the probability of absorption has a sharp rise. In the case of an isolated atom (blue line), a photo–electron is created and propagates away from the atom. When a neighboring atom is included in the picture (red line), NEXAFS occurs because the photo–electron can scatter back, modulating the amplitude of the photo–electron wavefunction at the absorbing atom. This in turn modulates the absorption coefficient, generating the fine structure.

Near Edge X–ray Absorption Fine Structure Spectroscopy (NEXAFS)

The region around the absorption edge contains information about both the electronic and the molecular structure. The low–lying excited states of a molecule are the ones that bind the systems together and contribute to determine their electronic structure [6]. The states populated by the excited

electrons in NEXAFS belong to this category: they encompass all the unoccupied states from the LUMO up to the ionization limit. In transition metals, the NEXAFS features include the unoccupied states of the narrow *d*–orbitals, just above the Fermi level, and the less tightly bound *s* and *p* bands in solids. For molecules or complexed ions, those states include unoccupied bound states as well as the low–lying continuum. Since these transitions are subject to the same selection rules as the optical ones, and since these rules can be relaxed by the local symmetry around the absorbing atom, information about the local geometrical arrangement can be obtained, for example, when forbidden transitions show up in the spectra. Other phenomena that contribute to shape the spectrum are multiple scattering resonances of molecules and condensed materials, causing modulations of the atomic X–ray absorption cross section of the absorbing atom.

X–ray techniques using absorption, offer the advantage of high penetration depth in matter. Recent developments could made use of quantitative NEXAFS to visualize the gradient of chemical composition across a section [11], rendering this technique capable of imaging chemical composition at sub–micron resolution. An active topic of research is the study of ultrafast phenomena relying on the pump–probe type of schemes [1, 2]. X–ray absorption can be used for not crystalline materials (amorphous or liquids), and it is atom–specific probe, being sensitive to the local environment.

Summarizing, NEXAFS is a powerful technique to investigate the nature of electronic structure and chemical bonding in condensed matter. It provides local probing, which translates into being able to detect nonequivalent environments, and it is sensitive to the unoccupied states character.

A large effort is conveyed to the assessment of the interpretation of spectra in order to understand how the character of intramolecular interaction affects the absorption, e.g. in relation to the coordination and bonding situation.The theoretical analysis often plays an accessory role to these assignments.

Molecular dynamics (MD) attempts to represent the effects exercised on a substance under given thermodynamic conditions, In this respect the behavior of the system is simulated by the means of an increased sampling via the microstates. MD simulations provide the sampling of the statistical ensemble, given specified thermodynamic contitions (T, V, p, μ), thus exploring

available configurations with the proper statistical probability. In the limit of a sufficiently extended sampling, physical properties, as spectra, can be averaged over the generated statistical configurations.

Calculating NEXAFS properties intrinsically needs a correct description of the core electrons, due to the fact that the quintessence of these phenomena lies on exciting an inner–shell electron to an unoccupied state or eventually to the continuum.

Several methods based on DFT have been developed to simulate the absorption spectra [10]. They can be grouped into two classes: in one case the core electrons are treated explicitly, alternatively, the PP approximation is employed and either a special core representation is needed or core molecular orbitals can be reconstructed a posteriori. In the former approach, normally due to its prohibitive computational cost, XAS calculations have been restricted to small samples or clusters (no periodic boundary conditions, or PBC), containing few non–hydrogen atoms and long range environmental effects cannot be properly taken into account. When PP are used, condensed matter simulations are possible, but, as mentioned, an *ad hoc* core–hole has to be approximated.

For this reason, relying on the recent implementation [5] of the GAPW method in `CP2K`, we want to perform DFT, linear scaling approaches which will possibly open the way to quantum chemistry prediction of XAS spectra of large molecules and extended condensed phase systems. Using `CP2K` the calculation of X–ray absorption spectra is carried out according to the Slater transition potential method (see next chapter for details). The GAPW formalism allows us to calculate the electronic structure including the core states also for condensed matter systems.

Chapter 3

Methodology of NEXAFS calculations

3.1 Slater Transition Potential Method

If an X–ray photon has sufficient energy to excite a core–level electron in a molecule, then the resultant photoelectron will leap into unoccupied states (Fig. (3.1)): this is the property that is explored by near edge X-ray absorption fine structure (NEXAFS).

The energy position of these transitions are unique to a given absorbing atom within a molecule, they contain spectral signals due to core–to–unoccupied valence orbital and core–to–continuum transitions. NEXAFS

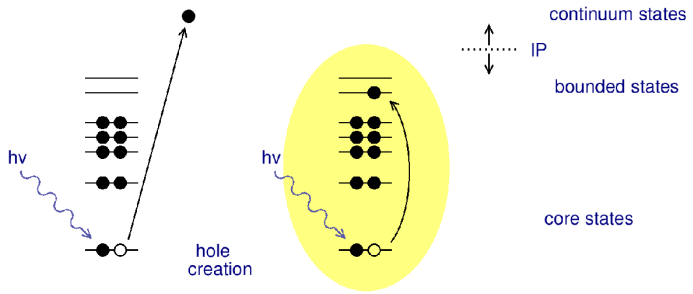


Figure 3.1: Illustration of a core–level ionization (left) and excitation processes (right).

contains thus information about both the electronic and molecular structure.

For this type of spectroscopy, the promotion of a core electron to an unoccupied orbital and the resulting relaxation of the electronic structure have to be appropriately modeled, and the electron–hole interaction effects too.

Introducing the occupation numbers $\{f_\alpha\}$ for the spin–orbitals Ψ_α, a solution of a parametric SCF calculation can be obtained, where the value of the energy in our context refers within Kohn–Sham theory (Eq. (1.7)):

$$\tilde{E}\big[\rho(\{f_\alpha\})\big] \qquad \rho(\mathbf{r}) = \sum_\alpha f_\alpha |\Psi_\alpha(\mathbf{r})|^2 \tag{3.1}$$

In principle, electronic transitions can be calculated as energy differences $\Delta E_{(i\to f)} = E_f - E_i$ for a transition between states i and f, instead of eigenvalues (ε) difference calculated for state i or f, $\Delta E_{(i\to f)} \neq \varepsilon_f - \varepsilon_i$.

This is known as ΔSCF and, in self–consistent field methods, it produces more accurate results since the energy difference includes explicitly the effects of relaxation of all the orbitals.

$$\begin{aligned} \mathrm{IP_i} &= \tilde{E}\left(0_i, 1_2, ..., 1_N, ..., 0_f, ...\right) - \tilde{E}\left(1_i, 1_2, ..., 1_N, ..., 0_f, ...\right) \\ \Delta\mathrm{SCF}_{(i\to f)} &= \tilde{E}\left(0_i, 1_2, ..., 1_N, ..., 1_f, ...\right) - \tilde{E}\left(1_i, 1_2, ..., 1_N, ..., 0_f, ...\right) \end{aligned} \tag{3.2}$$

However, this procedure is numerically not stable: this is the difference between 2 SCF calculations, which give rise to large cancellations. Furthermore, each state require to be fully optimized in the appropriate core–hole potential, and this scheme leads to difficulties with non–orthogonal and interacting MOs sets (Ψ_i and Ψ_f does not belong to the same subset of KS orbitals), making it not convenient to calculate a spectrum because each excitation necessitates one SCF convergence.

Following Slater's arguments on transition state theory, the energy difference can be obtained by the eigenvalue calculated at the occupation half–way between the two states [9], where the KS Hamiltonian has been modified by allowing a fractional change of the occupational number of i and f.

It is clear that for the $i \rightarrow f$ case the total energy is no longer a linear function of f_f. Let us assume, however, that the total energy can be expressed by inclusion of the next higher term as a second order polynomial in f_f. If we examine $\tilde{E}$ as a function of f_f, we see that the relaxed binding energy is given by the difference of the values of the curve at $f_f = 0$ and $f_i = 1$, or alternatively, by the slope of the chord connecting these two points of the curve. But for a parabola the slope of a secant between two points of the curve is equal to the slope of the curve itself at midpoint (in analogy with Cauchy mean value theorem):

$$\mathrm{IP_i} = \int_1^0 \frac{\partial \tilde{E}(\{f_j\})}{\partial f_i}\, df_i \simeq \left(\frac{\partial \tilde{E}(\{f_j\})}{\partial f_i} \right)_{f_i=1/2} = \varepsilon_i \left(\frac{1}{2} \right) \tag{3.3}$$

In contrast to the ΔSCF method, applying the Slater transition potential (TP), allows a direct calculation of excitation energies in one shot. Relaxation up to second order in $\partial \tilde{E}/\partial f$ are taken into account, and this approximation balances initial and final state contributions. One calculation has to be carried out for each atom, but with only one all electron SCF with the core–hole in the selected state i, the full XAS of this atom become available.

For example, an electron removal energy is the eigenvalue of the excited–core state computed when $\frac{1}{2}$ of an electron is missing in the given state, while a transition energy is the eigenvalue difference calculated when $\frac{1}{2}$ an electron is transferred between the two states.

As stated in Eq. (3.2), the electronic transition energies should be calculated by allowing the relaxation of the electrons. The major effect on the relaxation is exerted from the creation of the core–hole. On the contrary, the electron in the final state is immediately delocalized and does not affect the $\varepsilon_f^{\mathrm{TP}}$ and Ψ_f^{TP}. Adopting the half–core–hole (HCH) expression, the missing $\frac{1}{2}$ electron is set to the continuum, and the system finds itself $\frac{1}{2}$ positively charged. Thanks to this property, we can compute the entire spectrum for one specific atom by a single SCF cycle.

The relaxed orbitals are obtained in a single calculation by adjusting the conventional operators in KS in order to involve a fractional occupation number of the orbital of interest, and the transition energies are taken as difference between the resulting KS energies, $\varepsilon_f^{\mathrm{TP}} - \varepsilon_i^{\mathrm{TP}}$.

By treating all the system electrons explicitly, thus lifting the pseudopotential (PP) approach, there is no need of special core representation for the "ionized" atom, usually treated as an impurity by the introduction of an *ad hoc* PP, leading therefore to more accurate description of the excitation energies. Adopting this efficient approach, in the case of a solvated molecular species, this technique enables as application, an easier sampling along an MD run of the center characterized by different solvation shells.

In a solid or an extended system, it is not obvious how to carry out such a calculation, since there is no localized state whose occupation can be varied. One general approach is to identify such a localized state, e.g. a Wannier state or an approximation, and do calculations very much like those in an atom.

3.2 Spectra determination procedure

After a ground state Kohn–Sham calculation, few selected atomic core orbitals are projected on a minimal Slater basis $|\Psi_c(\mathbf{r} - \mathbf{R_A})\rangle$ the orbital to be excited is identified using maximum overlap criteria. The character of the core orbital, whether $1s$, $2s$ and so on, is determined by symmetry of the Slater orbital corresponding to the maximum overlap as given by the following expression.

$$\langle \Psi_c(\mathbf{r} - \mathbf{R_A}) \mid \Psi_\alpha(\mathbf{r}) \rangle \tag{3.4}$$

Once identified, the occupation number of the excited orbital is changed. The localized character of the core orbitals allows a simple atom–specific projection of the electronic structure.

As stated earlier, an unitary transformation is applied to the canonical orbitals, in order to obtain maximally localized orbitals, procedure that guarantees that the different core orbitals are disentangled and each one of them can be associated to one atomic center.

The nomenclature for XAS features reflects the core orbital from which the absorption originate. For example, K edges refer to transitions from the innermost $n = 1$ electron orbital, L edges refer to the $n = 2$ absorbing

electrons. The transitions are always referred to unoccupied states, *i.e.* to states with a photoelectron above the HOMO, leaving behind a core–hole, and absorption features may appear just below the edge, which correspond to transitions to bound unoccupied levels just below the ionization limit.

In the successive step we calculate the orbital energy differences by performing core–hole spin–polarized calculations (also known as unrestricted). Local spin density is necessary in order to account the explicit removal of the electron from the core.

The transition probabilities are governed by the dipole selection rules, and in the one electron picture they reduce to dipole moment elements between the initial and final orbitals. These integrals are calculated in the velocity form and averaged over the three Cartesian directions.

$$\hbar\omega_{if} = \varepsilon_f^{\mathrm{TP}} - \varepsilon_i^{\mathrm{TP}} \qquad \mathrm{I}_{if} = \frac{2}{3}\,\omega_{if}\,|\langle\Psi_i^{\mathrm{TP}} \mid \widehat{\mu} \mid \Psi_f^{\mathrm{TP}}\rangle|^2 \tag{3.5}$$

For the first transition, the TP method yields absolute transition energies whose value compare well to the more accurate ΔSCF estimations. A widespread procedure that is also here adopted consist to apply an overall shift of the entire spectrum obtained via TP based on the ΔSCF of the first transition computed based upon a fully relaxed core transition potential.

$$\delta_\omega = \omega_1 - \Delta\mathrm{SCF} \qquad \overline{\omega}_{if} = \omega_{if} - \delta_\omega \tag{3.6}$$

after this alignment the final spectrum is then absorption normalized (see Fig. (3.2).

To compare experimental and calculated NEXAFS transitions, we apply a Gaussian broadening of the spectral lines consisting of an increasing FWHM (represented by the exponent $\frac{1}{\sigma}$) ramp starting at the edge typically by an amount of 0.5 eV.

$$\sigma = \begin{cases} \sigma_{\min} & : \quad \omega_{if} < E_{\min} \\ \sigma_{\min} + \left(\omega_{if} - E_{\min}\right) \cdot \frac{\sigma_{\max} - \sigma_{\min}}{E_{\max} - E_{\min}} & : \quad E_{\min} < \omega_{if} < E_{\max} \\ \sigma_{\max} & : \quad E_{\max} < \omega_{if} \end{cases} \tag{3.7}$$

with typical values as:

$$\sigma_{\min} \sim 0.5 \text{ eV}, \quad \sigma_{\max} \sim 8 \text{ eV}, \quad E_{\max} - E_{\min} \sim 20 \text{ eV} \tag{3.8}$$

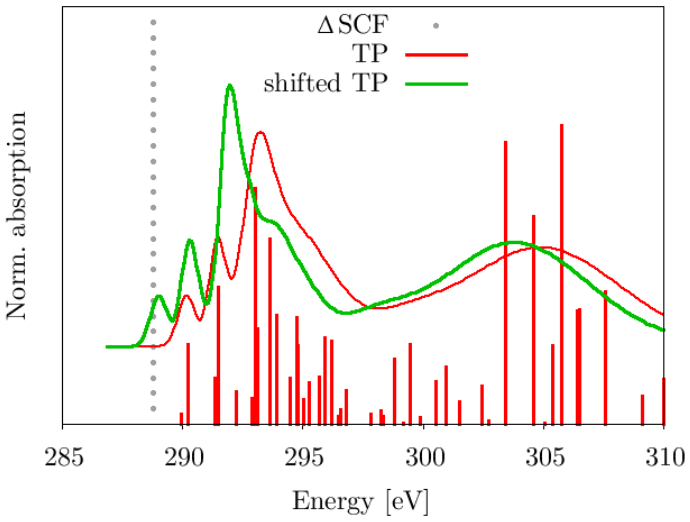


Figure 3.2: Example of a calculation: the spikes resulting from a TP calculation are convoluted with a superposition of Gaussians as described in Eq. (3.7) (red). The transitions are shifted to the ΔSCF value of the first transition (dotted line) by matching this value with the first TP resonance ($\sim$ 290 eV) by an alignment of δ_ω and the final spectrum is obtained (green).

References

[1] C. Bressler and M. Chergui. Ultrafast X–ray absorption spectroscopy. *Chem. Rev.*, 104:1781, 2004.

[2] M. Chergui and A.H. Zewail. Electron and X–Ray methods of ultrafast structural dynamics: advances and applications. *ChemPhysChem*, 10:28, 2009.

[3] F. de Groot. High–resolution X–ray emission and X–ray absorption spectroscopy. *Chem. Rev.*, 101:1779, 2001.

[4] A.P. Hitchcock. Inner shell excitation spectroscopy of molecules using inelastic electron scattering. *J. Electron. Spectrosc. Relat. Phenom.*, 112:9, 2000.

[5] M. Iannuzzi and J. Hutter. Inner–shell spectroscopy by the Gaussian and augmented plane wave method. *PhysChemChemPhys*, 9:1599, 2007.

[6] A. Nilsson and L.G.M. Pettersson. Chemical bonding on surfaces probed by X–ray emission spectroscopy and density functional theory. *Surface Science Reports*, 55:49, 2004.

[7] J.J. Rehr and R.C. Albers. Theoretical approaches to X–ray absorption fine structure. *Rev. Mod. Phys.*, 72:621, 2000.

[8] L.R. Sharpe, W.R. Heineman, and R.C. Elder. EXAFS spectroelectrochemistry. *Chem. Rev.*, 90:705, 1990.

[9] J.C. Slater and K.H. Johnson. Self–consistent–field Xα cluster method for polyatomic molecules and solids. *Phys. Rev. B*, 5:844, 1972.

[10] J. Stöhr. *NEXAFS spectroscopy*. Springer, 1996.

[11] S.G. Urquhart, A.P. Hitchcock, A.P. Smith, H.W. Ade, W. Lidy, E.G. Rightor, and G.E. Mitchell. NEXAFS spectromicroscopy of polymers: overview and quantitative analysis of polyurethane polymers. *J. Electron. Spectrosc. Relat. Phenom.*, 100:119, 1999.

Chapter 4

Carbon, Nitrogen, Oxygen and Sulfur NEXAFS Calculations of Peptides: Implications for the Building Block Approach in the Aqueous Environment

4.1 Introduction

In the last decade, NEXAFS spectroscopy has undergone a rapid development through the appearance of intense third–generation synchrotron radiation sources which have made selective excitation possible. The biochemically important C, N and O edges in the soft X–ray regime were not accessible with reasonable resolution until the development of high–resolution grazing incidence grating spectrometers, making soft X–ray spectroscopy increasingly popular for the characterization of biologically relevant systems.

Various detection schemes allow to tune the sampling depth, and therefore to study bioorganic samples in different environments ranging from ultra–high vacuum to solid [2, 11]. Within this context, Saykally and coworkers, have recently extended the domain of applicability of this technique pioneering both experiments and calculations in aqueous solution [23, 13]. NEXAFS

is sensitive to fine details of the electronic structure of organic molecules so that an analysis of the experimental data helps to reveal the character of the chemical bonding as well as small changes such as ones due to conformational transition or solvation, which are important, as example, for the metabolic functions of biomolecules.

Beside interests in exploring fundamental aspects of inner–shell excitation spectroscopy, a motivation for the present study is to support emerging efforts to differentiate and possibly map polypeptides in terms of conformation and tertiary structure in physiological conditions (i.e. in liquid water). Particularly, it is still a matter of investigation up to which extent a peptide spectrum is equivalent to a weighted sum of amino acid spectra, and how the difference in stacking interactions produces rationalizable spectral effects. In fact, together with theoretical simulations, NEXAFS results analysis is often performed in a simple picture based on the so–called "building block principle" (BB) [19], by which a molecule is seen as an assembly of smaller pieces. The total spectrum is thus divided into parts assigned to subunits (the building blocks) of the molecule that are identified by comparison with similar spectra of analogous molecules. In addition to accounting diatomic spectra as in the original definition, one can consider any functional group in an environment where it is probably little perturbed. However, there are clear limits to this procedure: delocalization of electronic charge across multiple functional groups leads to new molecular orbitals, combining the properties of the conjugated groups. In such cases the NEXAFS features of the individual groups can change significantly and new features may appear (breakdown of the BB).

4.2 Calculations

To calculate the spectra we rely on the density functional theory (DFT) approach and the Slater transition potential method. `CP2K` software package is an open source available implementation of these approaches [20], and was used for all molecular dynamic, energetic and spectral calculations presented in this work. In these calculations, we employed the BLYP form of the generalized gradient approximation to the exchange–correlation potential, along with 6–311++G** basis set for both the ground and the ionized states

for gaseous species and 6–311G** for condensed phase. An extensive analysis of the choice of parameters and the approaches required to calculate the NEXAFS spectrum in this context have been the subject of investigations described elsewhere [10].

Transition amplitudes are estimated in the single–particle and dipole approximations and excitations to states above this first excited state are approximated using the unoccupied Kohn–Sham eigenstates computed from the half–core–hole (HCH) self–consistent potential. Because of accounting for individual core electrons description, XAS calculations are carried out within an all–electron local spin density (unrestricted) framework.

The computed transitions are convoluted with Gaussian functions of reasonable line width in order to simulate the effect of both the limited photon resolution and the vibrational broadening of the bands in the discrete energy region and make the comparison with the experimental data easier. The width is chosen as a function of the term value: in the discrete region, the chosen line width is that of the instrument, while larger line widths are used for the lifetime–broadened continuum resonances. In comparing the computed and experimental spectra, the calculated transition potential energies are retained, and a rigid shift of the computed scale relative to a ΔSCF approach is applied.

In some studies, the core–excitation transition amplitudes are estimated without the impact of nuclear dynamics on the electronic subsystem. Often finite temperature effects are approximated by increased numerical broadening of calculated spectral peaks. A source of debate in the literature is related on how to incorporate vibrational effects explicitly in the calculation of NEXAFS properties. The approach of configurational sampling has been applied already in several XAS simulations, where distinct changes have been observed based on configurational modifications. As pointed out in the exemplar works of Uejio and Prendergast *et al.* [22, 17], it is important to incorporate motion in such calculations, whether sampling vibrational normal modes or averaging along a molecular dynamic (MD) trajectory. This is particularly true in a solvated environment context, due to the multitudes of solvation patterns around the molecular system of interest.

For a way out from this controversial aspect, we decide to adopt the sampling technique by the means of *ab initio* or classical MD simulations.

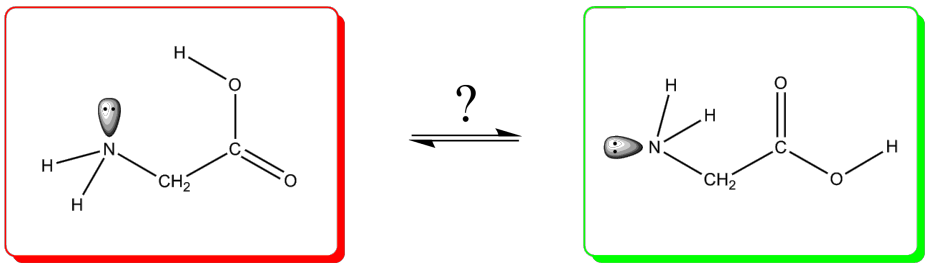
To perform a computer experiment the initial values for position and velocities have to be chosen together with an appropriate time step. The first part of a simulation is the equilibration phase in which strong fluctuation may occur. Once all important quantities are sufficiently equilibrated, the actual (or production) simulation is performed. Finally, observables are calculated from the trajectory.

4.3 Glycine

4.3.1 Isolated Systems

Glycine (Gly, NH_2–CH_2–COOH), also denoted α–aminoacetic acid, is the simplest amino acid. In the solid phase, glycine exists in form of a zwitterion, where the acidic hydrogen is transferred to the basic amino group. Upon heating, glycine vaporizes and converts to its nonionic form. In aqueous solution at neutral pH the zwitterionic form is dominant.

Several previous gas phase studies have been carried out to map gaseous Gly energetic landscape ([7] and references therein), attempting to find the possible existing conformers, however only a few amongst those candidates have been observed experimentally. Evidences provided by microwaves spectroscopy state that the population restricts mainly to two rotamers, namely Ip and IIp, where subscript p indicates a C_s symmetry (reflection symmetry only). Their structure is illustrated in Fig. (4.1).



(a) Conformer Ip **(b)** Conformer IIp

Figure 4.1: Amongst other possible conformers, Ip and IIp have been identified by microwave spectroscopy.

We turn our analysis towards these two plausible conformations in order to determine their influence upon NEXAFS signals, possibly also taking into account their respective contributions to the experimental reported spectra.

After geometry optimization a pure static calculation indicates an energetic preference towards Ip conformation respect to IIp in the order of few μeV. To investigate the sensitivity of inner–shell excitation spectra to the different glycine conformers at finite temperature, *ab initio* molecular dynamics trajectories are performed for both conformations. The gaseous specimens in the experiment are obtained by evaporation after heating, and the MD is conducted at high temperature accordingly, where 280 K were chosen. The calculations of the spectra are performed on selected snapshots of the trajectory with a sampling distance of 0.8 ps. Figure (4.2) illustrates the root mean square distance (RMSD) evolution of the structures along 8 ps of trajectory, this path is meant to generate sufficiently uncorrelated configurations for property calculation.

The subscript p for Ip and IIp in Fig. (4.1) drops out of the notation because the loss of intrinsic C_s symmetry due to molecular motion.

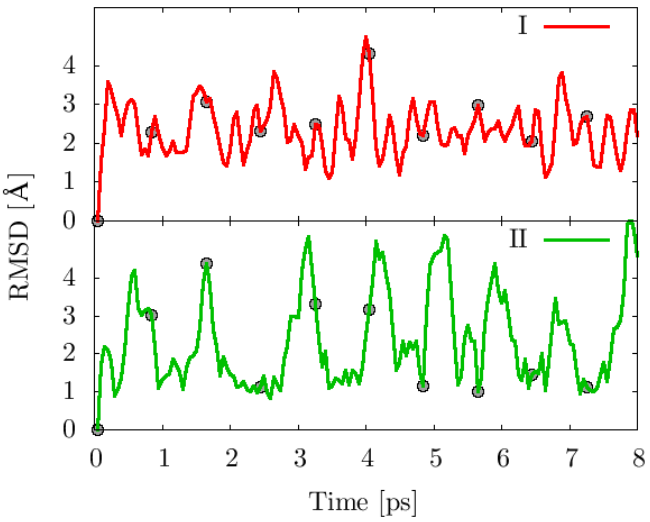
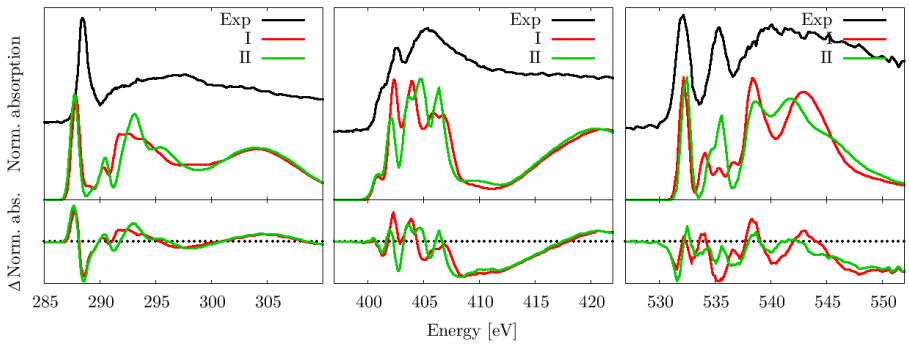


Figure 4.2: Positional root mean square displacement of conformer I (upper panel) and II (lower panel) as a function of simulation time. The gray circle points indicate the positions sampled.

Just like experimental conditions, where natural line widths are broadened, the resulting list of calculated vertical transitions and oscillator strengths are then convoluted with a series of Gaussians attempting to produce such an effect. Care was taken to maintain a reasonable width for each type of transition, varying between 0.8 eV for sharp $1s \rightarrow \pi^*$ transitions around the edge up to 8 eV for $1s \rightarrow \sigma^*$ and continuum transitions in the higher energy region. These values reflect also the transition lifetime. With this relatively small broadening we aim to simulate and distinguish electronic and vibrational effects explicitly, stressing that characteristic features could be easily put in evidence.

50 Kohn–Sham eigenstates are used to construct the transition matrix elements. This is sufficient to extend the spectra approximately to a window of 35 eV.

The comparison between the computed spectra of rotamer I and II are summarized in Figure (4.3). Experimental data are taken from the gas phase core excitation database collection maintained by the Hitchcock research group [9]. In particular we are referring to the high definition experiments of Gordon *et al.* [7].



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

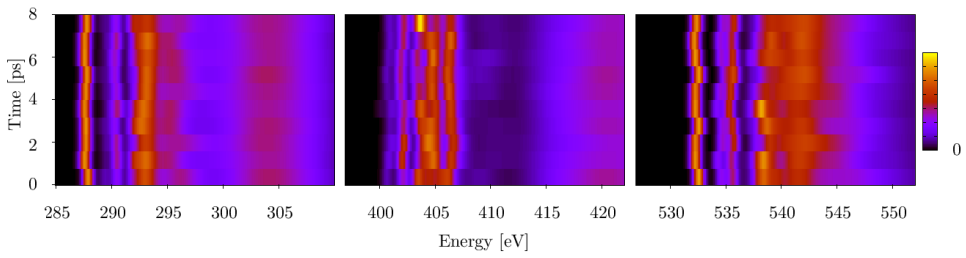
Figure 4.3: Gas phase glycine: in the upper panels are reported experimental NEXAFS (black line) and calculated average–spectra for the conformer I and II, vertical offsets are used for clarity for separate the curves. The lower panels illustrate the difference between theoretical and measured absorption for the two conformers compared with literature spectrum.

This graphs are realized by summation of the obtained signals for both the snapshots and the atomic species of a given atomic number, i.e. con-

former I oxygen spectra accumulates 20 calculations. The juxtaposition of the three graphs indicates the increasing energetic progression through the C 1s – N 1s – O 1s edges.

The transition attributions are performed by examining the spatial extent of the receiving unoccupied molecular orbital (not shown), their relative oscillator strengths and positions, and by comparison with the existing literature. Concerning C 1s edge (Fig. (4.3a)), both experimental and calculated I and II conformer spectra are dominated by the $s \rightarrow \pi^*_{\mathrm{C{=}O}}$ transition in the range of 287–288 eV. There is a $\sim$ 0.7 eV shift to lower energy for calculated transitions compared to the experimental one. The peaks centered at 290.5 eV are tentatively assigned to virtual valence levels of σ symmetry to be of $\sigma^*_{\mathrm{C–H}}$ and $\sigma^*_{\mathrm{C–N}}$, while the region 292–295 eV encompasses $\sigma^*_{\mathrm{C–OH}}$ and $\sigma^*_{\mathrm{C–C}}$. N 1s edge (Fig. (4.3b)) have relatively strong contributions from excitations to virtual levels of Rydberg character. The shoulders centered at 400.5 and 402 eV represents both transitions to $\sigma^*_{\mathrm{N–H}}$ orbitals, whereas in the range 404–406 eV are grouped $\sigma^*_{\mathrm{N–C}}$. O 1s edge (Fig. (4.3c)) shows two prominent discrete resonances: the two peaks of comparable intensity are characteristic of gaseous carboxylic acids and are associated with the existence of two oxygen atoms in very different environments, namely C=O versus C–OH. The peak at lower energy refers to $\pi^*_{\mathrm{C{=}O}}$ while the second to $\sigma^*_{\mathrm{C–OH}}$.

It is noteworthy to examine the evolution of the spectrum with conformational changes along the MD trajectory at 280 K.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.4: Gas phase glycine: conformer II (see Figure (4.1b)) 3D projection of the absorption as a function of simulation time. Every horizontal stripe representing a snapshot.

By explicitly displaying separately every frame contribution, Figure (4.4) indicates that along an equilibrium dynamics, molecular vibrations occurring in the gas phase do not affect dramatically the spectrum. In particular near the edge region, where the strong $s \rightarrow \pi^*$ transitions occur for the C and O atomic species, whereas soft modulations are induced in the $s \rightarrow \sigma^*$ and continuum excitations. The main peaks within 5 eV after the edge remain substantially unchanged as the trajectory progresses, while the position of the low intensity transitions at higher energies fluctuate. Indeed, those subtle changes contribute to shape the overall signal. Individual, single point calculations only based on the geometry of the optimized structure (Ip and IIp in Fig. (4.1)) do not allow a direct conclusion in terms of discriminatory sense. By accumulating several calculations it is possible to identify conformers that contribute more to the recorded spectra. The outcome of this procedure can be discussed on the basis of analyzing the deviation from the literature results compared by rotamer I versus II. The carbon spectra in Fig. (4.3a) for I and II are similar and not conclusive in this respect. The ones for nitrogen, depicted in Fig. (4.3b), display a first indication in that direction regarding the intensity relationship between the narrow peak situated at 402.5 and the one centered around 405 eV. In the experimental case the absorptions are 0.8:1 respectively. In structure I the same peaks intensities are 1:1 whereas for structure II are in a 0.75:1 ratio. The definitive hint comes by analyzing Fig. (4.3c) representing O 1s edge data. Experimental peaks centered at 532 and 535.5 eV indicate a 1.1:1 relative intensity: in both conformers the first peak, coming from the C=O moiety, is reproduced with the same intensity, the second peak, related to the C–OH part, is not pronounced for conformer I, while present in II, although slightly underestimated in comparison with the value of the literature.

The conjunction of these information indicate that a better match between experiments and calculations is obtained for conformer II. Comparison with the intensities ratio between the features suggest a major role of geometry type II which accounts of a predominant contribution to the overall spectrum. Focusing on conformer I oxygen spectrum, the weak intensity of the C–OH signal compared to the C=O could never give rise to the observed characteristic experimental peaks. The calculations support the idea that conformer II has the highest relative abundance at the estimated temperature of the gas in the experiments.

From this type of considerations emerges that, in principle, it is possible with NEXAFS to investigate how the nature of a given conformation of a compound influences the spectrum. Computational results attempt to help in the interpretation of observed signals in terms of different possible configurations giving information on which states are more probable to contribute to the experimental spectra.

4.3.2 Solid Phase

While isolated glycine molecules in the gas phase are neutral, intermolecular interactions occurring in the condensed state favor proton transfer from the carboxy group to the amino group and give rise to the glycine zwitterion ($^{+}NH_3$–CH_2–COO^{-}).

Solid glycine is known to exist in three polymorph modifications, α–, β– and γ–glycine, which differ from each other in the pattern of hydrogen–bonds formed by interacting with near glycine zwitterions. The thermodynamically most stable crystal structure is the α phase. In α–glycine each molecule forms three intermolecular NH$\cdots$O bonds with neighboring molecules. Figure (4.5) displays the H–bond (HB) network present in the structure investigated.

Table 4.1: Solid glycine: quantitative description of the H–bonds involved. For the localization of the HB patterns see Figure (4.5).

NH$\cdots$O	distance [Å]	angle [deg]
HB*	2.85	163
HB**	2.77	168
HB***	3.07	155

In the present study, large periodic crystals are investigated. The α form structure is taken from the crystallographic structure deposited at the Cambridge Crystallographic Data Center (CSD entry `glycin19`) [3]. The experimental values of the monoclinic primitive unit–cell dimensions are used for calculations, scaled up to 10.2 x 11.9 x 10.9 Å^3, with respectively 90, 111.8, and 90 angle degrees, in order to accommodate 16 glycine molecules.

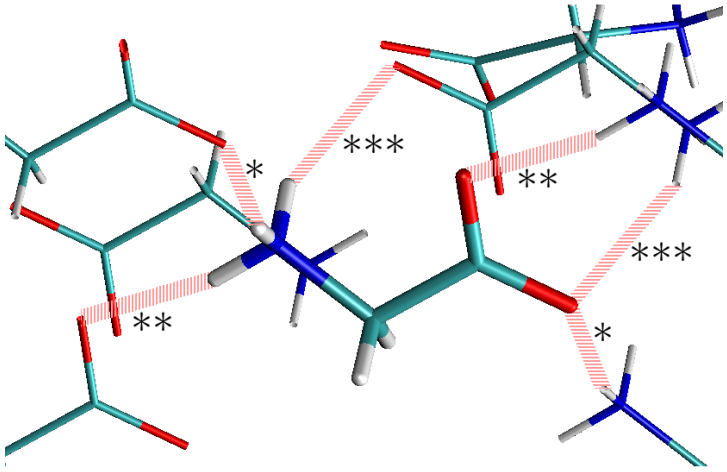
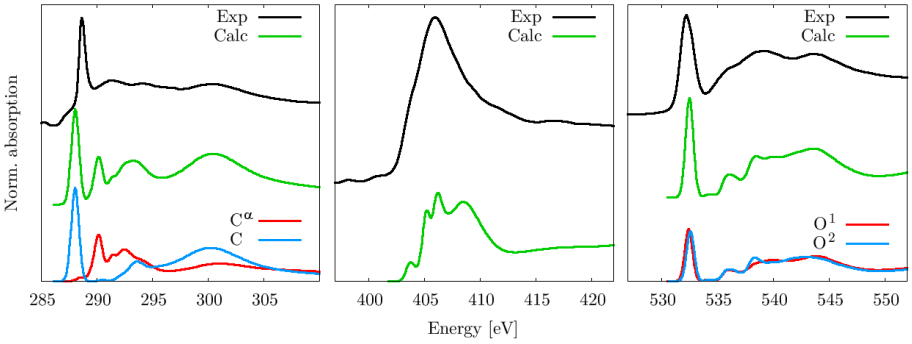


Figure 4.5: Solid glycine: experimental structure of α–glycine. Magenta color lines indicates intermolecular NH$\cdots$O hydrogen–bond (HB) network. O^1 is involved in HB**, while O^2 is participating in HB* and HB***. See Table (4.1) for HB's quantitative details.

In solid Gly, due to its intrinsic electronically compact phase, 500 Kohn–Sham states are necessary to cover a 35 eV excitation window.

The results of a single structure calculations of the 16 molecules are depicted in Figure (4.6), where the individual atomic contributions to the total spectrum are explicitly plotted.

The calculated C $1s \rightarrow \pi^*_{C=O}$ transition at the edge is like in the gas phase shifted 0.7 eV towards lower energies with respect to the experiment. The peaks assignment follows the progression like in gas phase, except for the missing $\sigma^*_{C–OH}$ transitions. In the N $1s$ spectrum of α–glycine the pre–edge shoulder characteristic of glycine in gas phase is quenched. This feature associated with $\sigma^*_{N–H}$ transitions is normally substantially attenuated in condensed phase, due to the charge transfer associated with hydrogen bonding. Similar effects have been observed also in water, where the H_2O $\sigma^*_{O–H}$ transitions are present in gas phase, but vanish in the spectrum of ice. An alternative explanation from Messer *et al.* [13] is that the difference arises from the fact that the nitrogen belongs in one case in a neutral ($–NH_2$) versus cationic ($–NH_3^+$) moiety in solid state, thus giving rise to two distinct sub–



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.6: Solid α–glycine: incremental offset are applied. From the top to the bottom are depicted experimental [24] (black lines) and calculated (green lines) NEXAFS absorptions. The lower curves represent the single atomic species. In the carbon spectra, the notation C^{α} indicate the $–CH_2–$ atom while C is related to $–CO_2^-$.

species. This interpretation is supported by the pH dependency of solvated amino acids in which this behavior is reflected (for a detailed discussion see Section 4.3.3). Solid specimen exclusively exhibit a single low–lying $\pi^*_{C=O}$ resonance at the O 1*s* edge. The two distinct features of gaseous glycine (Fig. (4.3c)) coalesce into a single peak, with a transition energy close to the lower energy peak, consistently with the expected zwitterionic form, where only the carboxylate environment is present. The higher intensity of the peak with respect to the continuum in Gly(s) relative to Gly(g), is indeed a consequence of both oxygen in Gly(s) atoms being in the same environment.

Overall, glycine solid phase simulation results show a very structured contour. In contrast to the gas phase calculations, in the solid case no sampling procedure is applied. The introduction of libration effects implied by the temperature would probably cure on this respect by influencing the lateral motion of the energetic scale of the peaks, just upon the same phenomena encountered in Figure (4.4). Finite temperature effects on nitrogen spectrum have been the subject of other studies [16].

As introduced before, the two nominally equal oxygen atoms present in the chemical formula of Gly(s) find themselves in a slightly different H–bond environment (see Fig. (4.5) and Tab. (4.1)), O^1 being involved in one HB while O^2 in two, although one normal and one weak HB. Calculations

indicate that albeit this formal distinction, it is not sufficient to differentiate the spectra, based only on the number and the strength of HB's. Similarly, a substantial impact in the individual signals in the lower part of Figure (4.6c) is not observed, but for a slight perturbation located at 538.5 eV quantifiable in an intensity surplus of 10 % compared to the main peak absorption for O^2 versus O^1. No additional information are acquired by analyzing the molecular orbitals of the final states (not shown) involved in the transition generating this light profile discrepancy.

4.3.3 Aqueous Phase

In aqueous solutions, three charge states of glycine, namely zwitterion, glycinium cation and glycinate anion, coexist in equilibrium, with their mutual ratio being strongly dependent on the pH of the solution. In aqueous solution at low pH, glycine is found in its cationic form, at neutral pH the zwitterionic specie is dominant, and at high pH the anion is the predominant form. More precisely, the speciation for the zwitterion is in an acidity regime of $2.34 < \text{pH} < 9.6$.

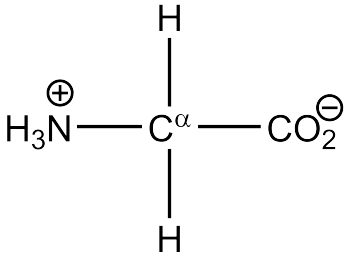


Figure 4.7: Zwitterionic glycine.

From the experimental point of view, recent advances in the field of NEXAFS spectroscopy, especially the ones pioneered by the team of Saykally, have made possible to extend the principles of this techniques also to liquid water as a medium, unraveling thus the possibility to perform new investigations compatible to physiological conditions. This technological and conceptual breakthrough has been achieved by coupling small jets of water solution to a synchrotron beamline: usually a liquid microjet of ~ 30 μm diameter stream is injected into vacuum and windowlessly intersected with the X–ray beam.

In the experiment of aqueous phase glycine [13], a 0.6 M solution sample is used and the temperature has been determined to be 288–293 K when probed by the X–ray beam, recording C, N and O $1s$ edges.

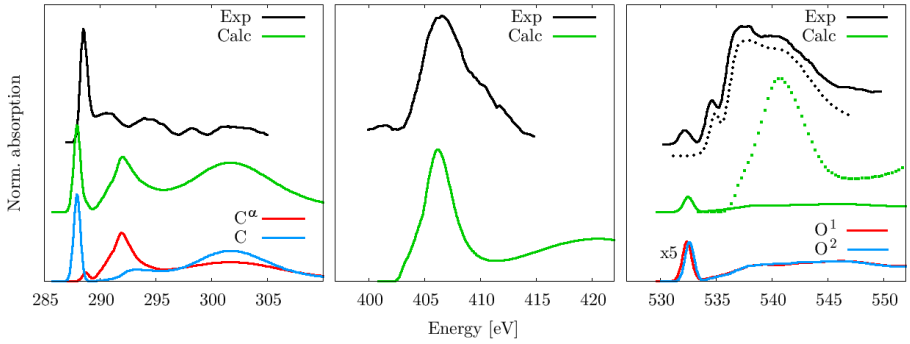
Like the previous gas phase calculations, based on the strategy to sample configurations along a MD trajectory, we aim to simulate the behavior of a solvated glycine (Gly(aq)), where in addition to the vibrations of the molecule of interest, the complexity of having to deal with a surrounding solvent is added. Water is a polar molecule, forming hydrogen–bonds (HB), both between other water molecules and with polar moieties of the solute. At finite temperature the solvation shell around the solute is fluxional and undergoes substantial rearrangement during time. In order to be able to sample the slow rearrangement of the solvation shell we employ classical MD simulation, which allows for large simulation times and size.

Relying on Amber force field parameters for the solute [4], the MD simulations were performed in an approximately $(12\ \text{Å})^3$ orthorhombic simulation cell under periodic boundary conditions (PBC) contouring solute and 89 TIP3P water molecules.

A production run of 350 ps is generated according to the NVE ensemble. The resulting distribution of nuclear coordinates is sampled at regular intervals of 3.5 ps, thus allowing 100 snapshots.

NEXAFS simulations of Gly(aq) employed the same PBC conditions as the MD simulations, explicitly including all waters in the quantum mechanical description.

Figure (4.8a) summarizes the results for the carbon edge. Although the carbon spectra of gas, solid and aqueous species are quite similar, there are noticeable differences in the width and relative intensities of the resonances located between 290 and 300 eV. A precise spectral assignment C $1s \rightarrow \pi^*_{\text{C=O}}$ is only possible for the sharp resonance observed near 289 eV, in spite of the large spectral density modulations between 290 and 305 eV. No single transition in this region appears to dominate the spectrum. The broad resonances can be associated to transitions to a variety of σ^* and Rydberg states. For nitrogen and oxygen edges there is little qualitative difference between the solid and aqueous spectra, but both appear substantially distinct from that of the gas phase. Although it is generally believed that



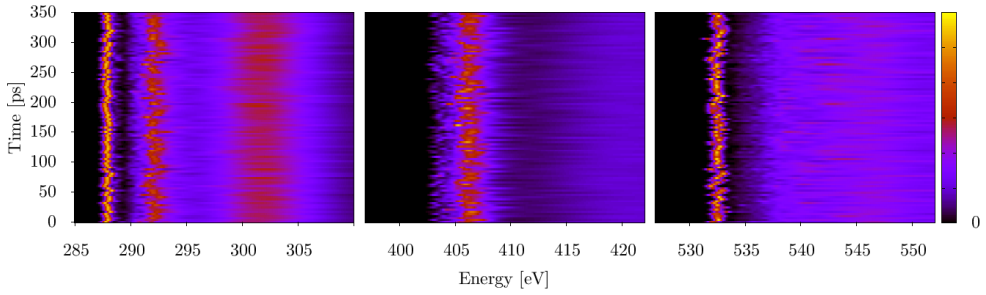
(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.8: Aqueous glycine: vertical offset are applied. Experimental (black lines) and calculated (green lines) NEXAFS absorptions. The lower curves represent the single atomic species. In the oxygen spectra, the dotted lines report neat water, while O^1 and O^2 stand for the chemically equivalent carboxylate oxygens, where the absorption has been magnified by a factor of 5.

Rydberg transitions broaden into a continuum upon condensation, this is not the only possible explanation for the N 1*s* spectral differences observed between gas and solvated Gly. On the basis of pH dependence of the electronic structure of glycine, Messer *et al.*[13] suggested that they can be associated to distinct $-NH_2$ versus $-NH_3$ moieties. This thesis is confirmed by another pH–dependent study, namely aqueous lysine photoelectron spectra (XPS) [14], where the sequential protonation of the two amino groups present in this molecule suggests the possibility to monitor quantitative distinct environments generated by charge densities modifications induced by selective proton attachment. Other studies supporting this argument are based on glycine oligomers solid phase XPS spectra [5].

In the experimental spectra in Figure (4.8c), the O 1*s* spectra for Gly(aq) (black solid line) and neat bulk water (black dotted line) are illustrated as a comparison. The results of the calculations on Gly(aq) are separated into the component coming from the two oxygen of the carboxylate of glycine (green solid line) and the oxygen coming from the surroundings solvent molecules (green dotted line). By inspection of the two experimental spectra, it turns out that the spectra of Gly(aq) differs mainly in the peak located at 542 eV. It is reasonable to argue that this peak (arising from glycine) falls out-

side the water region. The calculated spectrum has been divided into two contributions, according to the chemical identity of the species, namely the signal derived from oxygen of the carboxylate moiety and those associated to water molecules. The purpose of this partition is to verify that also computationally one does not observe a spectral overlap between these two distinct species, thus avoiding some superposition phenomena that would induce artifacts in the rationalization of the results. The spectra of O^1 and O^2 superimpose: this indicates that the chemical equivalence condition for the two oxygen is fulfilled via the adopted sampling method. The same shape of the signals indicates that sampling over sufficiently large number of configurations, the two oxygens become equivalent, i.e. experience the same solvation conditions.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.9: Aqueous glycine: 3D projection of the absorption as a function of simulation time. Every horizontal stripe representing a snapshot separated by 3.5 ps apart.

Inclusion of MD sampling results in a clear extra broadening (Fig. (4.9)) induced by small displacements of the nuclei. The modulations of the peaks in the edge and near edge region are quantitatively more pronounced than in the gas phase, this phenomenon can be linked to the combination of both vibrational effects and to those induced by the variety of hydration environments.The importance of including the explicit description of the solvating environment in calculations has been addressed in many ways, and in NEXAFS context has been put in evidence by Uejio [21].

In this work, we found that using 100 snapshots is sufficient to achieve

convergence. Surprisingly, by following a given atomic spectrum along the trajectory, quenching or attenuation effects are observed. Sometime a particular peak separates in two contributions. This can be rationalized in a combination of both vibrational effects of the solute and partial charge transfer between glycine and surrounding waters. However we could not find a clear correlation between a given solvation pattern and specific spectral features. This is probably beyond the capabilities of a single atom consideration, meaning that it is really a concerted phenomenon involving the inferring of the hydrogen–bond network with the solvent that extends beyond the first sphere of solvation. Increased complexity is added if one considers that together with HB's involving neighboring water molecules, occasional forms of "bifurcated" bond with two adjacent water are observed in the MD trajectory. The combination of those factors make out of reach such a rationalization based simply by mapping the nature of the HB and the structure of the spectrum.

4.4 Peptides

4.4.1 Glycine Oligomers

GlyGly

The simplest specie in the biomolecular context which incorporates a peptide bond is glycyl–glycine (GlyGly, $^{+}NH_3$–CH_2–CONH–CH_2–COO^{-}), in which two glycine molecules are combined in a head–to–tail arrangement with elimination of a water molecule.

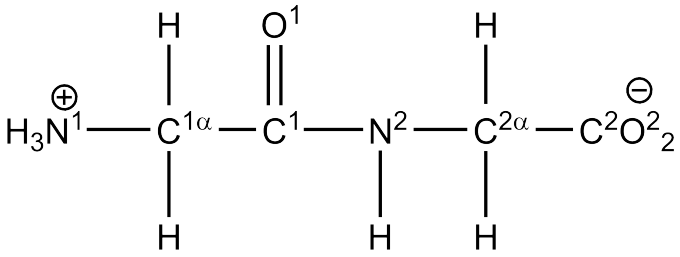
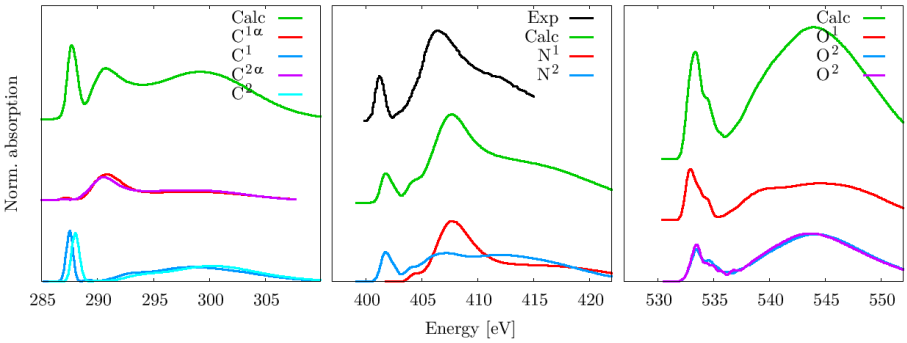


Figure 4.10: Zwitterionic diglycine.

In order to simulate GlyGly(aq), we use a box of 117 water molecules. This correspoinds to a more diluted solution with respect to what we used for the monomer, to account for the smaller solubility of the dimer.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.11: GlyGly(aq): incremental offset are applied. From the top to the bottom are depicted experimental (black lines) and calculated (green lines) NEXAFS absorptions. The lower curves represent the single atomic species. For the atoms nomenclature see Fig. (4.10).

The NEXAFS spectra extracted from the MD trajectory of the dimer are than compared to those obtained for the monomer.

When comparing the outcome of Gly(aq) and GlyGly(aq) C 1*s* calculations, there is a 0.3 eV shift to lower energy and a broadening between the two spectra. The differences can be attributed to the fact that the monomer contains one simple carboxylate group, while for the dimer the contribution of the carboxylate and those of the amide are mixed. A similar shift can be observed in a previous comparison of the carbon experimental spectra in terms of change from a carboxyl to an amide group in both gas and solid phase. The origin of that shift in terms of change from a carboxyl to an amide group has been also discussed [2]. In most experimental C dipeptide solid spectra, it can be seen that the peak resulting from the C=O bond near 289 eV is shifted also about 0.3 eV towards lower energies in comparison to glycine [7], which is clearly an effect of the peptide bond. The other peak position of the σ^* resonances after the main edge are unchanged. By analyzing GlyGly(aq) N 1*s* edge in comparison to Gly(aq) a new pre–edge feature is observed: the strong resonances near 402 eV (feature 1) comprise near–degenerate transitions localized along an NH single bond in the amide

nitrogen. This feature is not present in the nitrogen spectrum of Gly, since the absence of peptide bond. The peak centered around 407 eV (feature 2) encompasses σ^* resonances like in Gly(aq). It is important to notice that those transitions are broadened for GlyGly(aq) in comparison to those of Gly(aq). This broadening could be due to additional contributions from weaker transitions in this energy range or from inhomogeneous broadening that does not occur with glycine, suggesting a variety of solvation environment. The broad shoulder after the two main peaks, before the continuum region, is shifted further to higher energies in the diglycine spectrum than in glycine where the signal rapidly decays. The shoulder is therefore either due to a different set of transitions or the ${\sigma^*}_{\mathrm{NH}}$ transitions have experienced significant environmental broadening. By averaging 100 MD snapshots, the adopted approach leads to an intensity ratio between features 1 and 2 of 0.35, comparable to the experimental value of 0.5 [12], and improving towards a smoother continuum region above 410 eV. O 1s spectra of GlyGly(aq) exhibits a single peak at 533.5 eV, which is dominate by the ${\pi^*}_{\mathrm{CONH}}$ transition but has minor contributions from ${\pi^*}_{\mathrm{COO^-}}$ transitions.

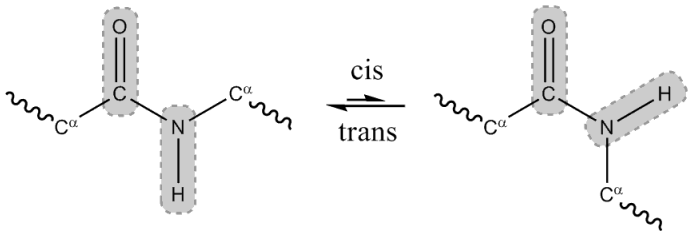
GlyGly Cis–Trans Isomers

The carbon–nitrogen bond constituting the peptide bond has a partial double bond character. A possible test of conformational–dependence of the NEXAFS profile is related to the difference in geometry between cis–trans isomers around the peptide bond (Fig. (4.12)). From an energetic point of view, the trans form is overwhelmingly adopted in most peptide bonds, however, mainly because of intramolecular strain or long–range interactions, some peptide incorporate the cis form.

This structural distinction could in principle be reflected on the overall atomic signal, providing thus insights into the electronic structure changes that accompany the two different peptide bonds orientations.

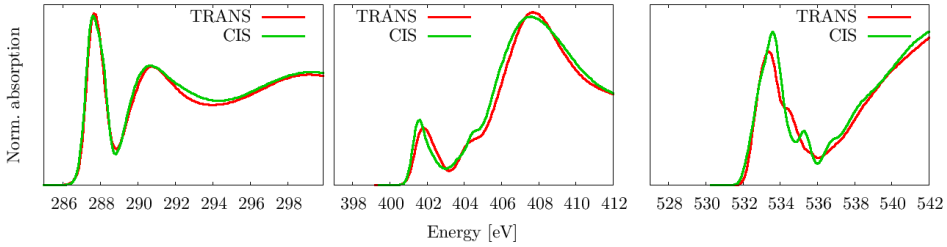
In this case it is deliberately chosen to narrow the Gaussian broadening to better expose distinct spectral features underlying the entire spectrum.

By analyzing the cis and the trans signals in Figure (4.13), no particular



(a) Conformer trans **(b)** Conformer cis

Figure 4.12: Trans–cis isomerization: highlight at the peptide bond.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.13: Trans and cis GlyGly (aq): superposition of cis and trans conformers NEXAFS signals.

features can be pinpointed. This a priori promising interesting source of distinction between two topologies of the amide linkage apparently does not give rise to characteristic patterns in the signal that would allow possible discrimination between the two states.

Building Block Fragments

A key tenet within NEXAFS analysis and interpretation of the data is the role played from molecular additivity concept [19]. An example of such principle for the rationalization of biomolecular solid phase results has been introduced in the form of a modified building block approach, the "X–ray absorption spectral simulator" [18], a code that generates a sum based on the amino acid sequence and then adds a distortion which mimics the spectral modifications arising from the structural changes associated with peptide bond formation. Hereafter is proposed a scheme that involves the contri-

butions of individual units. Substitutional groups have been determined to explore their fingerprinting character using the building block (BB) concept in the context of peptides. In principle the peptide bond spectrum is a differential correction to an amino acid which incorporates changes associated with the loss of the amine and carboxylic groups and the addition of the amide group.

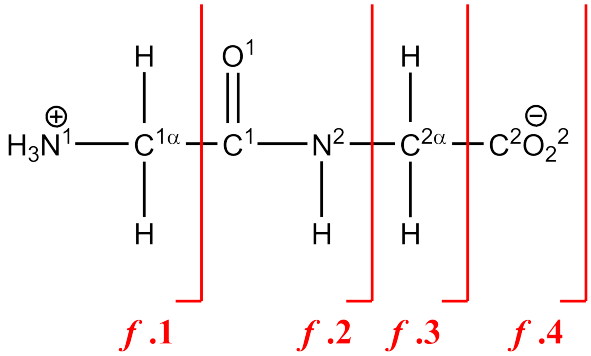


Figure 4.14: Building block (BB) fragments obtained from GlyGly(aq): from left to right, $f.1$ incorporates atoms in the amino terminal end, $f.2$ in the amide bond, $f.3$ in the alpha carbon and $f.4$ within the carboxyl–terminus.

In Figure (4.14) is depicted the subdivision into structural motifs used as BB for the aqueous phase. Combining those fragments any homopeptide of glycine can be built.

GlyGlyGly

Two amino acids form a dipeptide by condensation and polypeptides are created by repetition of this process. The tripeptide diglycyl–glycine (GlyGlyGly $^{+}NH_3$–CH_2–$(CONH–CH_2)_2$–COO^{-}) is obtained combining three glycine as monomer unit.

To test the validity of molecular additivity based on the fragments illustrated in Fig.(4.14), for the case of GlyGlyGly the total signal is obtained as the following function:

$$\mathrm{BB}\{\mathrm{GlyGlyGly(aq)}\} = \{f.1\} + 2\times\{f.2\} + 2\times\{f.3\} + \{f.4\} \quad (4.1)$$

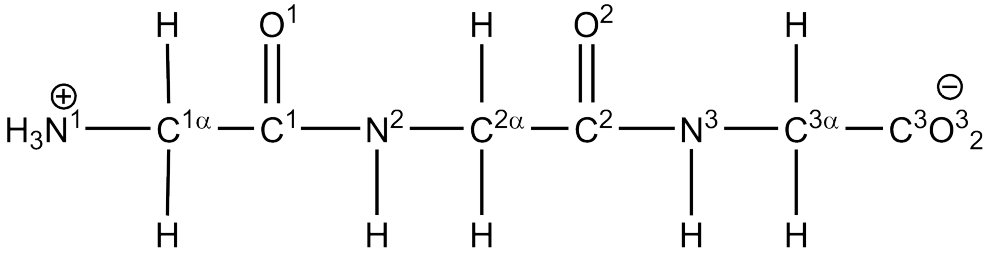
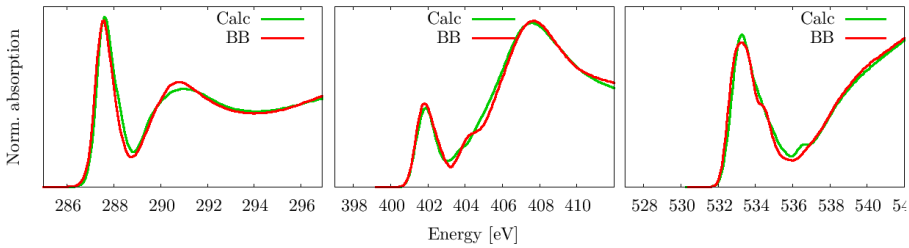


Figure 4.15: Zwitterionic triglycine.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.16: GlyGlyGly (aq): superimposition of standard MD trajectory–calculated signal (Calc) and obtained by applying the building block (BB) principle.

Diglycine and triglycine have carbamido groups with the ratios of carboxyl to carbamido groups equal to 1:1 and 1:2 respectively.

Since most nitrogen atoms in a peptide are amide nitrogen sites, the N 1s spectrum of GlyGlyGly should be similar to the N 1s spectra of GlyGly: this feature is observed and we note that the π^* amide signal is actually much stronger than that in GlyGly as expected.

The two spectra in Figure (4.16) overlap at low energies, validating the reliability of the choice of BB fragments and their relative calculated absorption profiles. This decomposition illustrate how efficiently an aqueous phase NEXAFS signal can be casted in therms of unitary components.

Like in the cis–trans case, where the influence of the geometrical conformation have been inspected to check if it would have had an impact on the signal, it is worth to investigate the influence of the peptidic ϕ and ψ dihedral angles. These two dihedral angles describe the angles involved in the pep-

tide bond moiety: ϕ describes the rotation around C^n–N–C^α–C^{n+1} bonds, while ψ the rotation of the N^n–C^α–C–N^{n+1} bonds, where the superscript n indicates the index of the residue. An illustration of the ϕ and ψ dihedral characterizing mainchain distribution is sketched in Figure (4.17).

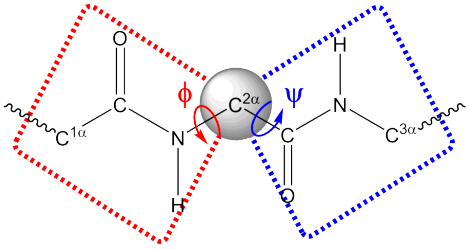
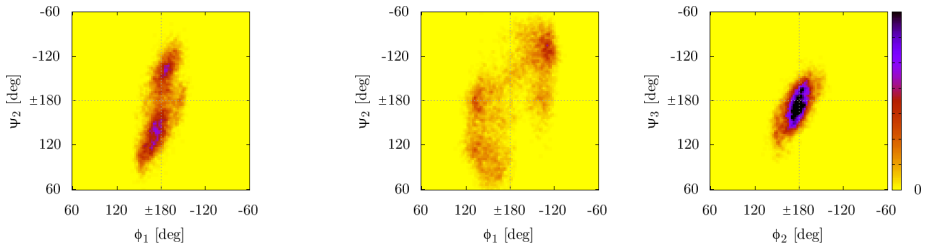


Figure 4.17: ϕ–ψ schematic representation of planes characterizing peptide bond mainchain dihedral.

Figure (4.18) displays the distribution of dihedral angles sampled along a 350 ps MD trajectory.



(a) GlyGly **(b)** GlyGlyGly

Figure 4.18: Ramachandran plot: ϕ–ψ dihedral angles for GlyGly(aq) vs GlyGlyGly(aq) of a 350 ps MD trajectory.

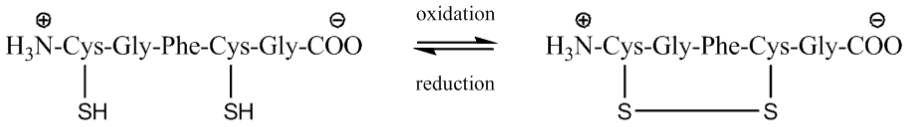
The Ramachandran plot of GlyGly(aq) in Fig. (4.18a) indicates two major zones of distribution, while the first set of ϕ–ψ dihedral angle for GlyGlyGly(aq) almost uniform distribution over a broader area. Contrarily to this latter smearing, the second indicates a well centered, compact conformational preference in one peak.

In spite of the fact that the dihedral angle distribution is significally different for GlyGlyGly(aq), meaning that the sampling configuration space is broader for the trimer, the spectral features are still similar to those calculated with GlyGly(aq) as a basis.

This outcome suggests that the molecular rotation around the dihedral has rather negligible effect on the absorption spectra, in all three core edges, and there is no apparent sensitivity towards cis or trans peptide bond.

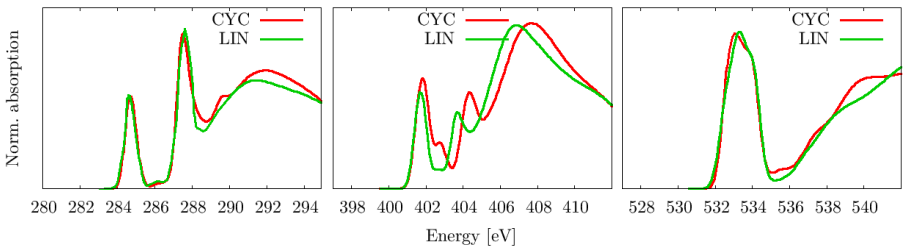
4.4.2 Sulfur Containing Polypeptide

A polypeptide is a relatively short sequence of amino acid residues joined into a unique molecule via the peptide bonds. Some polypeptides contain disulfide bonds: these are cross–links between chains or between parts of a chain, formed by the oxidation of cysteine residues (Fig. (4.19)). The presence of such bonds has an impact on the three–dimensional structure, because the the –S–S– bridge would stabilize certain geometries by covalent linkage. In a peptide with the sequence Cys–Gly–Phe–Cys–Gly, the oxidized structure forms a small loop (Fig. (4.19b)), whilst in the reduced one there is no defined structure.



(a) Linear conformer **(b)** Cyclic conformer

Figure 4.19: Disulfide bond: a disulfide bridge (–S–S–) is formed from the thiol groups (–S–H) of two cysteine residues. The product of the oxydation is a cystine residue.



(a) Carbon **(b)** Nitrogen **(c)** Oxygen

Figure 4.20: Cyclic and linear peptide: aqueous spectra for both oxidized and reduced form of the pentapeptide illustrated in Fig. (4.19).

Carbon edge spectra presented in Fig. (4.20a) are substantially different from previous C 1s calculated ones. Indeed, although the x–coordinates window frame extends over the usual 25 eV, its abscissa value is shifted by 5 eV towards lower energies. This in order to accommodate the low–lying π^* antibonding states resonances coming from the aromatic ring of phenylalanine. The presence of this residue shapes the overall spectra by introducing additional transitions, thus shifting the NEXAFS edge by -2.8 eV, in agreement with literature results [2]. The N 1s edge calculations displays a somewhat pronounced additional $\sigma^*_{\text{N–CH}_2}$resonances, localized around 404 eV.

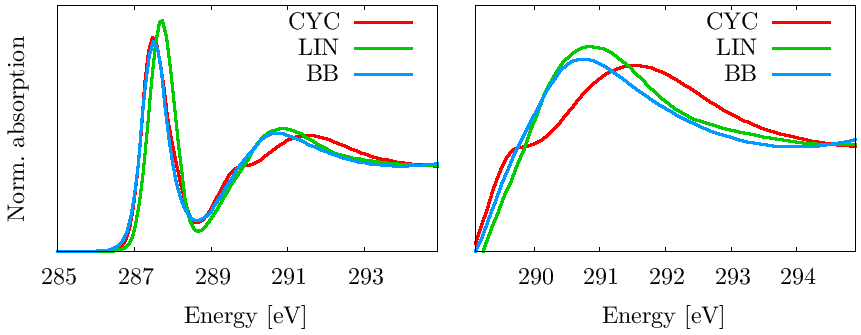
Considering the overall C, N and O calculated signals, it is not obvious how to discriminate between the cyclic and the linear form of the pentapeptide, since the spectral profiles show very similar features.

The BB reconstruction of the C 1s edge spectral profile based on GlyGly(aq) (4.14) calculations semms not to be adequate for this peptide. Among the possible reasons, one is for sure related to the presence of additional atoms in the lateral chain of the amino acids of the peptide, making not directly compatible the two topologies. Regardless, a conciliating comparison can be obtained by considering only the atoms present in the peptide mainchain. Thus the obtained BB reconstruction refers to the backbone carbons. In this respect it holds:

$$\text{BB}\{\text{main chain(aq)}\} \;=\; \{f.1\} \;+\; 4\times\{f.2\} \;+\; 4\times\{f.3\} \;+\; \{f.4\} \quad (4.2)$$

In Figure (4.21), we report the spectral profile computed for the backbone C atoms only. The computed spectra of linear and cyclic peptide are compared to the BB reconstructed, where as basis GlyGly(aq) have been used. The three curves are in good agreement, in particular at lower energies. Some discrepancy is revealed by magnificating the region between 289 to 295 eV, where the band of the cyclic spectrum is shifted to higher energies. The BB profile follows more closely the linear peptide behaviour, as expected.

A key point of the building block model is that the spectral signature of a functional group does not change from the reference to the final chemical compound, or if there is a change, it is uniform and well–quantified.



(a) C 1*s* mainchain **(b)** Zoom into C 1*s* mainchain

Figure 4.21: Backbone C 1*s* cyclic and linear peptide(aq): main chain reconstruction from the cyclic and the linear pentapeptide calculated from MD trajectories and via the BB approach.

This agrees with the fact that BB contributions have been parametrized via GlyGly(aq), molecule that does display a linear conformation, without any perturbation coming from elements of secondary structure stabilized by intramolecular H–bonds or disulfide bridge like it would be the case with the cyclic form. This remarkably good agreement between the linear form and the BB–obtained spectrum displays how transferable are in principle the contributions starting from the simple fragmentation proposed in Fig. (4.14). The problem, however, is that the small difference between cyclic and linear, that would allow in principle to be used for differentiation purposes, will find itself smeared when adding all the other carbon contributions coming from the lateral chains. The BB profiles for N and O edge overlap in the NEXAFS region: only at higher energies they diverge and fall somewhere in between the linear and the cyclic form (data not shown).

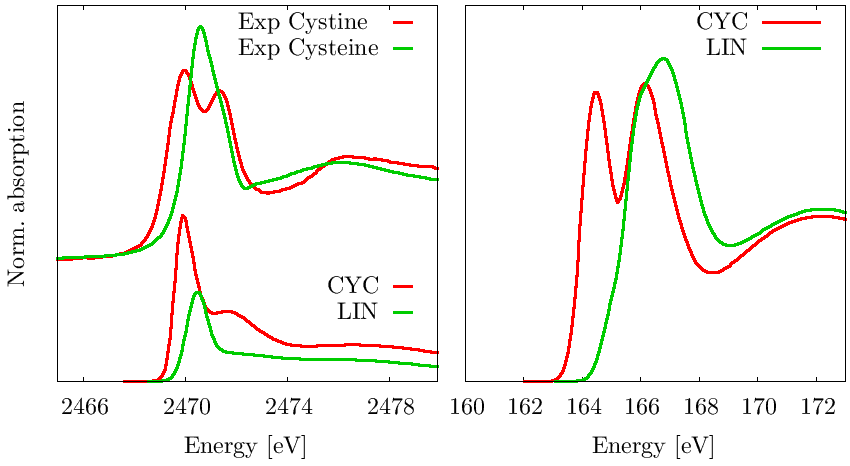
Again, the calculations indicate that, although we are considering nominally different atoms, equivalent chemically but slightly different in terms of local geometrical parameters in the conformationally labile peptide molecule, no special spectral changes can be individuated.

Together with protein–based experimental studies, the outcome of this calculations indicate that diversity is nearly absent in protein spectra, due

to averaging over a multitude of residues [18, 26]. In smaller peptides, or proteins with relatively unusual sequences, is the contribution of individual amino acids or classes of amino acids that can be observed [25].Therefore, there are no indications that C, N, and O NEXAFS analysis alone are able to differentiate between the reduced and the oxidized form of a disulphide–prone peptide. In the case of a cysteine–containing peptide there is although another atomic specie whose edge could be investigated: namely the sulfur atom. In contrast to the large abundance of the aforementioned C, N and O elements in proteins, where sensitivity to different environment is weak and contribution is hidden in the overall spectrum, the presence of only two S atoms offers the advantage that its signal is univoque and directly linked to the source of singularity between the two structure, because this element is involved in the disulphide bridge construction. This makes the S edge particularly appealing since it is a single probe element within the structures.

Recent studies of Sandström *et al.* [15], focus on S 1s NEXAFS in an aqueous context both for different biochemically relevant molecules as well as their spectral dependence upon pH change. Although the calculations we present involves the sulfur atom embedded in a protein environment, a fair quantitative comparison can be obtained, keeping in mind that the measured results refer to simple amino acids.

In contrast to C, N, and O 1s edges, electrons within the S 1s orbitals can experience the influence of relativistic effects. The simple ΔSCF scheme to align the spectra needs a further correction to take into account this type of contribution. In their experimental paper, Sandström and coworkers attempt to calculate the overall contribution to be able to use the aforementioned alignment procedure. This would be a shift towards lower energies for the ΔSCF part to whom have to be added the positive term shifts due to the relativistic effects, making an overall shift of ~ 4 eV towards higher energies. This rigorous approach is nevertheless abandoned here in favor to a more pragmatic alignment to experimental results. The linear form of the peptide, incorporating cysteine, is compared with neat cysteine, and the cyclic form, containing the disulfide bond, is compared with neat cystine. By aligning the calculated S 1s resonances to the experimental edge peak an overall shift of 0.8 eV is applied.



(a) S 1*s* edge **(b)** S 2*s* edge

Figure 4.22: Cyclic and linear (aq) spectra: two different S edges and two different peptide spectra are presented. Literature measurements refers to solvated neat cysteine and cystine [15].

The qualitative features of the spectra in Fig. (4.22a) are similar. Namely two peaks for the cyclic form compared to cystine and one for linear compared to cysteine. Concerning the intensity ratio between the peaks in the cyclic form, the second peak results sensibly lowered and broadened, and in the linear form it appears more like a shoulder at higher energies. For the linear form the character of the main transitions is $\sigma^*_{\text{S–H}}$, whereas the shoulder comes from $\sigma^*_{\text{S–C}}$ excitations. The cyclic form main peak is related to $\sigma^*_{\text{S–S}}$, while the second peak to $\sigma^*_{\text{S–C}}$.

Based on the results of this calculations, there is potentially a 0.5 eV difference, together with an intensity variation, between the first peak in the two forms of the peptide, namely the reduced linear and the oxidized cyclic conformation. In principle the sensitivity of the usual detectors, lower bounded to 0.1 or 0.2 eV, would permit to differentiate amongst this structures. However a more striking difference in the shape of the spectrum is obtained by comparing the S 2*s* edge. The 2*s* edge, although not ΔSCF corrected, indicates that the two peaks for the cyclic form are of equal intensity, and the shoulder in the linear peptide is incorporated in the main peak. Furthermore the two edges are 1.5 eV apart, possibly making spectral identification easier,

due to the distinct profile displayed from the structures. The $2s$ NEXAFS receiving unoccupied molecular orbitals are similar in nature to the $1s$, with the notable difference in the larger intensity of the relative peaks, in contrast to the broadening. The calculated energetic range for the $2s$ edge of the two forms of the peptide corroborate with the energetic range reported in the literature for a thiophene compound [1], suggesting that a relativistic corrections to $2s$ electrons could play a minor role in this case.

4.5 Conclusions

We presented NEXAFS DFT calculations and analysis of a set of biochemically relevant substances: glycine in its gaseous and solid phase, as well as aqueous glycine and its series of solvated oligomers. Calculations are in good agreement with experiments, further validating the proposed theoretical approach. A second set of simulations have been performed on a pentapeptide susceptible to undergo cyclic to linear transition, mediated by the creation of disulphide bridge via two cysteine residues. C, N and O $1s$ edges and S $1s$ and $2s$ edges have been calculated for the two structures.

The spectra are collected over large sampling of the configurational space, as obtained by MD at 300 K. As expected the spectral signal are affected by variations in the chemical environment of each individual atom. In this respect, sampling vibrations and different conformations are necessary to reproduce the broadening of spectral features observed in experiment. The overall signal depends on entropic factors, i.e. due to the coexistence of a large number of chemically equivalent atoms within an individual molecule, which, however, experience slightly different intramolecular (as defined by the exact molecular conformation) and intermolecular (H–bonding, etc.) environments.

This poses also a challenge in using this kind of spectroscopy for fine structural refinement, when there are many almost equivalent atoms experiencing slightly different chemical environment. Such situation produces broad experimental spectra, where distinctive feature characterizing the system are hard to be found.

Qualitatively, it is possible to correlate specific NEXAFS features with functional groups and in some cases, individual bonds, such that the total

spectrum can be considered as a linear combination of elementary spectra. This conceptual approach, named the "building block principle", provides a useful starting point for the interpretation of the spectra of very complicated molecules.

The C, N and O calculations show no particular evidence for spectral contrast for potential protein–protein identification. Because of the large numbers of amino acids in a polypeptide, there is a high degree of spectral averaging, making proteins with "typical" ratios of amino acid residues almost indistinguishable. Differentiation among proteins is a quite difficult task and relies on spectral differences caused by difference in amino acid composition or, potentially, structure. One exception to this indistinguishability by averaging, could be proteins composed of a repeating motif, such as many structural proteins [18], However, the X–ray absorption spectra of proteins are quite distinct from that of other biological materials. This has been the basis for mapping locations and concentrations of proteins relative to other biological materials or carbon–based polymers, where experiments show chemical speciation at better than 50 nm spatial resolution contrast [8].

According to sulfur NEXAFS calculations, S $1s$ edge spectroscopy does not allow a straightforward identification between sulfhydryl and disulfide bonds, whilst S $2s$ edge is able to better capture the related modifications in the electronic structure.

This feature could be used to identify structural arrangement associated to the formation of disulfide radical anions. Such reaction can be induced by electron uptake in proteins, occurring after radiation–triggered damages [6], conventionally studied via classical pump–probe type experiments.

Hence, NEXAFS analysis can become useful in distinguishing conformational changes of biosystems, when clearly identifiable probe elements are present. Further studies in this direction should be considered, for example addressing the influence given by the presence of prosthetic groups or cofactors, like a porphirine ring or metal ions. Another possibility would arise by the role played by phosphorylation, enzymatic mechanism that adds a phosphate (PO_4) group to a protein, process often related to cellular signal transduction.

Acknowledgments

Experimental data for the gaseous phase glycine were made available by A.P. Hitchcock via the "Gas Phase Core Excitation Database", while the solid α–glycine ones by M. Zharnikov. Access to aqueous phase data were kindly provided by C. Cappa (C, N and O edges) and M. Sandström (S edge).

References

[1] J. Birgerson, M. Keil, Y. Luo, S. Svensson, H. Ågren, and W.R. Salaneck. A study of the electronic structure of ethylenedioxythiophene in gas phase using NEXAFS and quantum chemical calculations. *Chem. Phys. Lett.*, 392:100, 2004.

[2] J. Boese, A. Osanna, C. Jacobsen, and J. Kirz. Carbon edge XANES spectroscopy of amino acids and peptides. *J. Electron. Spectrosc. Relat. Phenom.*, 85:9, 1997.

[3] Cambridge Crystallographic Database Center. http://www.ccdc.cam.ac.uk, 2013.

[4] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, et al. AMBER 9. *University of California, San Francisco*, 2006.

[5] A. Chatterjee, L. Zhao, L. Zhang, D. Pradhan, X. Zhou, and K.T. Leung. Core–level electronic structure of solid–phase glycine, glycyl–glycine, diglycyl–glycine, and polyglycine: X–ray photoemission analysis and Hartree–Fock calculations of their zwitterions. *J. Chem. Phys.*, 129:105104, 2008.

[6] E. Dumont, A. D. Laurent, and X. Assfeld. Intersulfur distance is a key factor in tuning disulfide radical anion vertical uv–visible absorption. *J. Chem. Phys. Lett.*, 1:581, 2010.

[7] M.L. Gordon, G. Cooper, C. Morin, T. Araki, C.C. Turci, K. Kaznatcheev, and A.P. Hitchcock. Inner–shell excitation spectroscopy of the peptide bond: comparison of the C 1s, N 1s, and O 1s spectra of glycine, glycyl–glycine, and glycyl–glycyl–glycine. *J. Phys. Chem. A*, 107:6144, 2003.

[8] A.P. Hitchcock, C. Morin, X. Zhang, T. Araki, J. Dynes, H. Stöver, J. Brash, J.R. Lawrence, and G.G. Leppard. Soft X–ray spectromicroscopy of biological and synthetic polymer systems. *J. Electron. Spectrosc. Relat. Phenom.*, 144:259, 2005.

[9] Hitchcock Research Group. http://unicorn.mcmaster.ca/corex/cedb-title.html, 2010.

[10] M. Iannuzzi and J. Hutter. Inner–shell spectroscopy by the Gaussian and augmented plane wave method. *PhysChemChemPhys*, 9:1599, 2007.

[11] K. Kaznacheyev, A. Osanna, C. Jacobsen, O. Plashkevych, O. Vahtras, H. Agren, V. Carravetta, and AP Hitchcock. Innershell absorption spectroscopy of amino acids. *J. Phys. Chem. A*, 106(13):3153–3168, 2002.

[12] B.M. Messer, C.D. Cappa, J.D. Smith, W.S. Drisdell, C.P. Schwartz, R.C. Cohen, and R.J. Saykally. Local hydration environments of amino acids and dipeptides studied by X–ray spectroscopy of liquid microjets. *J. Chem. Phys. B*, 109:21640, 2005.

[13] B.M. Messer, C.D. Cappa, J.D. Smith, K.R. Wilson, M.K. Gilles, R.C. Cohen, and R.J. Saykally. pH dependence of the electronic structure of glycine. *J. Phys. Chem. B*, 109:5375, 2005.

[14] D. Nolting, E.F. Aziz, N. Ottosson, M. Faubel, I.V. Hertel, and B. Winter. ph–induced protonation of lysine in aqueous solution causes chemical shifts in X–ray photoelectron spectroscopy. *J. Am. Chem. Soc.*

[15] E.D. Risberg, F. Jalilehvand, B.O. Leung, L.G.M. Pettersson, and M. Sandström. Theoretical and experimental sulfur K–edge X–ray absorption spectroscopic study of cysteine, cystine, homocysteine, penicillamine, methionine and methionine sulfoxide. *Dalton Trans.*, 2009:3542, 2009.

[16] C. P. Schwartz, R. J. Saykally, and D. Prendergast. An analysis of the NEXAFS spectra of a molecular crystal: alpha–glycine. *J. Chem. Phys.*, 133:44507, 2010.

[17] C. P. Schwartz, J. S. Uejio, R. J. Saykally, and D. Prendergast. On the importance of nuclear quantum motions in near edge X–ray absorption fine structure spectroscopy of molecules. *J. Chem. Phys.*, 130:184109, 2009.

[18] J. Stewart-Ornstein, A.P. Hitchcock, D.H. Cruz, P. Henklein, J. Overhage, K. Hilpert, J.D. Hale, and R.E.W. Hancock. Using intrinsic X–ray absorption spectral differences to identify and map peptides and proteins. *J. Phys. Chem. B*, 111:7691, 2007.

[19] J. Stöhr. *NEXAFS spectroscopy.* Springer, 1996.

[20] The CP2K developers group. http://cp2k.berlios.de/, 2010.

[21] J.S. Uejio, C.P. Schwartz, A.M. Duffin, A. England, D. Prendergast, and R.J. Saykally. Monopeptide versus monopeptoid: insights on structure and hydration of aqueous alanine and sarcosine via X–ray absorption spectroscopy. *J. Chem. Phys. B*, page 685, 2010.

[22] J.S. Uejio, C.P. Schwartz, R.J. Saykally, and D. Prendergast. Effects of vibrational motion on core–level spectra of prototype organic molecules. *Chem. Phys. Lett.*, 467:195, 2008.

[23] K.R. Wilson, M. Cavalleri, B.S. Rude, R.D. Schaller, A. Nilsson, L.G.M. Pettersson, N. Goldman, T. Catalano, J.D. Bozek, and R.J. Saykally. Characterization of hydrogen bond acceptor molecules at the water surface using near–edge X–ray absorption fine–structure spectroscopy and density functional theory. *J. Phys. Cond. Mat.*, 14:L221, 2002.

[24] Y. Zubavichus, A. Shaporenko, M. Grunze, and M. Zharnikov. Solid–state near–edge X–ray absorption fine structure spectra of glycine in various charge states. *J. Phys. Chem. B*, 110:3420, 2006.

[25] Y. Zubavichus, A. Shaporenko, M. Grunze, and M. Zharnikov. NEXAFS spectroscopy of homopolypeptides at all relevant absorption edges: polyisoleucine, polytyrosine, and polyhistidine. *Journal of Physical Chemistry B*, 111:11866, 2007.

[26] Y. Zubavichus, A. Shaporenko, M. Grunze, and M. Zharnikov. Is X–ray absorption spectroscopy sensitive to the amino acid composition of functional proteins? *J. Phys. Chem. B*, 112:4478, 2008.

Part III

Nuclear Magnetic Resonance

Chapter 5

Magnetic Linear Response Properties Calculations With the GAPW Method

Nuclear magnetic resonance (NMR) and electron paramagnetic resonance (EPR) are powerful spectroscopic techniques, providing invaluable insights in the atomic structure of materials across a broad range of scientific disciplines. In recent years, there has been a growing interest in the *ab initio* quantum mechanical calculation of the quantities extracted from NMR/EPR spectra within Kohn–Sham density functional theory (DFT) [16, 23]. A comprehensive overview of the various approaches in this field is given in Ref. [18]. Comparing experimental NMR/EPR quantities with those computed from proposed molecular models, would open the possibility to derive structure/spectroscopic correlations from e.g. molecular dynamics. This would provide a basis for determining numerous aspects of the molecular structure and its related properties, such as chemical bonding and chemical reactions, which are not readily accessible from experiment.

Many interesting scientific problems that would potentially benefit from a theoretical NMR/EPR study involve simulations that easily require many thousands of atoms, such as nanostructures, interfaces, molecular liquids, and complex biomolecules in their natural environment. Thus, there is still the need of developing efficient algorithms for the calculations of such properties for systems containing many thousands of atoms. Recently, Ochsenfeld

et al [35] and Kussmann and Ochsenfeld [26] introduced a method for the computation of chemical shifts for gas phase systems that scales linearly with the number of atoms. The authors report a large Hartree–Fock NMR calculation of N–methyl nicotinamide surrounded by water molecules with a total of 1003 atoms.

In this work we apply a method for the calculation of NMR chemical shifts based on the Gaussian and augmented–plane–wave (GAPW) [31, 25, 17] formalism in its all–electron (AE) version. A key step in the calculation of the aforementioned magnetic properties is the capability to compute the induced current density generated by the external static magnetic perturbation. The present approach relies on density function perturbation theory (DFPT) [1, 13, 12, 39].

A previous implementation of the DFPT, in a planewave framework, has been successfully applied for the calculation of the chemical shift [42, 36, 43] as well as to the g tensor [7] for condensed systems. However, since the original technique makes use of the planewave plus pseudopotential representation of the electronic structure, its applications have been restricted to the lightest element (hydrogen). The proposed method differs from the work by Sebastiani and Parrinello [42], and Mauri *et al* [34] in a few respects. It represents the electronic structure with local atomic centered Gaussian functions allowing for reduced complexity algorithms. Thus large scale calculations of the NMR and EPR parameters become feasible. The all electron description of the system permits the evaluations of the chemical shifts and g tensor for all elements.

The development of the DFPT within the GAPW formalism allows an all electron description of the induced current density, thus lifting the main drawback of the planewave implementations [42, 34]. To this purpose, it is necessary to extend the concepts of the GAPW representation of the electronic density to the current density induced by the external magnetic field.

To our knowledge, the gauge–including projector augmented wave (GIPAW) method [37, 38] is the only other AE method (using a frozen–core approach) currently able to calculate these quantities for condensed phase systems using periodic boundary conditions (PBC).

The use of localized orbitals for the evaluation of the magnetic properties

within PBC was originally introduced by Sebastiani and Parrinello.

Among all the different methods available in the literature (see e.g. [18]), the gauge–including atomic orbital (GIAO) introduced by London [32] and first adopted for quantum chemical calculation of the NMR parameters by Ditchfield [8] is known to be the best for evaluating NMR shifts [18, 22]. In spite of the superiority of the GIAO approach, we chose the individual gauge for atoms in molecules (IGAIM) approach introduced by Keith and Bader [19] and the continuous set of gauge transformation (CSGT) approach by the same authors [20]. This choice was driven by the simplicity of the implementation of the IGAIM and CSGT compared to the GIAO approach.

We can also note that the implementation of the hyperfine coupling tensor has been recently presented within the GAPW formalism [6].

The structure of this chapter is as follows: first, the theoretical aspects of the method are elaborated, with particular attention on features that are specifically related to the GAPW representation. This includes the decomposition of the current density in soft and hard terms, the convergence with respect to basis set size and the choice of the gauge. Then, the accuracy of the approach is presented by comparing the results obtained for a set of small isolated molecules to the values obtained from well–established methods commonly used for NMR calculations. Finally, one illustrative application is presented: a quantum mechanical/molecular mechanical (QM/MM) calculations of the NMR shifts of an adenine molecule hydrated with 827 water molecules. In this example, the QM part contains 66 atoms.

5.1 Theory

The $xy-$components of the chemical shift tensor σ^A corresponding to nucleus A, for systems with net electronic spin 1/2, can be evaluated through the following expressions (note that atomic units will be adopted throughout)

$$\sigma^A_{xy} = \frac{1}{c}\int_{\Omega_S}\left[\frac{\mathbf{r}-\mathbf{A}}{|\mathbf{r}-\mathbf{A}|^3}\times\mathbf{j}_x(\mathbf{r})\right]_y d^3r \tag{5.1}$$

where c is the speed of light in vacuum, $\mathbf{A}$ is the position of the nucleus A, $\mathbf{j}_x$ the current density induced by a constant external magnetic field applied

along the x axis, and Ω_S is the volume of the whole integration domain, including the periodic replicas of the simulation cell. The other tensor components can be obtained by changing the indices accordingly.

$\mathbf{B}_x^{\mathrm{corr}}$ represents the magnetic field that originates from the corresponding total induced current density $\mathbf{j}_x$, and is given by

$$\mathbf{B}_x^{\mathrm{corr}}(\mathbf{r}) = \frac{1}{c}\int_{\Omega_\mathrm{S}} \frac{\mathbf{r}'-\mathbf{r}}{|\mathbf{r}'-\mathbf{r}|^3} \times \left(\mathbf{j}_x(\mathbf{r}') - \mathbf{j}_x^s(\mathbf{r}')\right) d^3r' \tag{5.2}$$

where the subtraction of $\mathbf{j}_x^s = \mathbf{j}_x^\alpha - \mathbf{j}_x^\beta$ is introduced as a self–interaction correction.

It is readily apparent from Eqs. (5.1 – 5.2) that the induced current densities are a key ingredient in the evaluation of the NMR properties. In the following sections we describe first how the induced current densities can be derived in DFPT as applied within the GAPW formalism, and then we turn our attention on the actual evaluation of the tensor components.

5.2 Calculation of the Induced Current Densities

The linear response of the electronic structure due to the application of an external magnetic field can be determined by solving a system of linear equations of the type

$$-\mathrm{i}\sum_{il}\left(H_{kl}\delta_{ij} - S_{kl}\int \psi_i^{(0)}(\mathbf{r})H(\mathbf{r})\psi_j^{(0)}(\mathbf{r})d^3r\right)C_{li}^{(1)} = \sum_l H_{kl(j)}^{(1)}C_{lj}^{(0)} \tag{5.3}$$

where the index i runs over the occupied ground state orbitals, H_{kl} is a matrix element of the unperturbed Hamiltonian, S_{kl} is an element of the overlap matrix, $H_{kl}^{(1)}$ is a matrix element of the perturbation operator, and the matrix $C^{(1)}$ is the matrix of the expansion coefficients of the corresponding linear response of the orbitals, $\psi^{(1)}$. Note that the imaginary nature of response orbital has been made explicit, thus allowing us to work with real expansion coefficients also for the response orbitals. The optional subindex (j), labeling

the matrix element of the perturbation operator, indicates that the perturbation might be orbital–dependent. In the case of interest here, the perturbation can be split into three different types of operators, which are $(\mathbf{r}-\mathbf{d}_j)\times\mathbf{p}$, the *orbital angular momentum* operator (which leads to the response orbitals C^L), $\mathbf{p}$ the *momentum* operator (leading to C^p), and $(\mathbf{d}_i - \mathbf{d}_j) \times \mathbf{p}$ the *full correction* operator (leading to $C^{\Delta i}$). In the notation, the vector $\mathbf{d}_j$ is the Wannier center associated with the unperturbed j–th orbital. It makes the angular momentum and the full correction perturbation operators dependent on the unperturbed orbital to which they are applied. In conclusion, the linear response orbitals are then given by 9 sets of expansion coefficients, C^L, C^p and $C^{\Delta i}$, as for each operator the three Cartesian components are considered. The matrix elements of the perturbation operator are

$$H^L_{klj} = -\mathrm{i}\int \chi_k(\mathbf{r})(\mathbf{r}-\mathbf{d}_j)\times\nabla\chi_l(\mathbf{r})d^3r, \tag{5.4}$$

$$H^p_{kl} = -\mathrm{i}\int \chi_k(\mathbf{r})\nabla\chi_l(\mathbf{r})d^3r, \text{ and} \tag{5.5}$$

$$H^{\Delta i}_{klj} = -\mathrm{i}(\mathbf{d}_i-\mathbf{d}_j)\times\int \chi_k(\mathbf{r})\nabla\chi_l(\mathbf{r})d^3r. \tag{5.6}$$

We can note here that for a magnetic field as a perturbation, the first order change in the charge density vanishes everywhere in space. Thus this perturbation does not give rise to a first order change in the Hartree and exchange–correlation terms. This simplifies considerably the linear system of equations and is often called uncoupled perturbed self–consistent field equations.

Once all the contributions to the linear response orbitals have been calculated, the x–component of the linear current density response induced from an external magnetic field applied along the y axis can be written as

$$\begin{aligned} j_{xy}(\mathbf{r}) = -\frac{1}{2c}\sum_{ikl}\Big[C^{(0)}_{ki}\Big(C^{L_y}_{li} + (d(\mathbf{r})-\mathbf{d}_i)_x C^{p_z}_{li} - (d(\mathbf{r})-\mathbf{d}_i)_z C^{p_x}_{li} - C^{\Delta i_y}_{li}\Big) \\ \times\{(\nabla_x\chi_k(\mathbf{r}))\chi_l(\mathbf{r}) - \chi_k(\mathbf{r})\nabla_x\chi_l(\mathbf{r})\}\Big] \\ + (\mathbf{r}-d(\mathbf{r}))_z\rho(\mathbf{r}) \end{aligned} \tag{5.7}$$

where $d(\mathbf{r})$ is a gauge that shall be discussed in the following sections of this work. While the first term (square brackets), on the right–hand side of Eq. (5.7), represents the paramagnetic contribution to the current density, the second term is the diamagnetic part. The latter vanishes identically when the CSGT approach [20] is employed, *i.e.* $d(\mathbf{r}) = \mathbf{r}$. The other components of the current density are obtained in analogous way by changing appropriately the Cartesian indices.

5.2.1 The Position Operator in PBC

The position operator $\mathbf{r}$ operating on a (one–particle) wave function in the coordinate representation $\psi(\mathbf{r})$ results in the multiplication of this wave function with the position variable $\mathbf{r}$. When periodic boundary conditions are imposed, the multiplicative position operator is not a valid operator, since the Cartesian components of $\mathbf{r}\,\psi(\mathbf{r})$ are no longer periodic [40]. In the derivation of the response orbitals, the position operator appears in the definition of the perturbation operators and in the definition of the current density. To solve this problem, we maximally localize the ground state orbitals [49, 33, 3]. For insulators, these Wannier functions feature an exponential decay [4] and can be considered as confined for sufficiently large simulation cells. As described in Ref. [42], the position operator can be re–defined to obey the PBC by using a sawtooth–shaped profile centered at the Wannier center of the localized orbital to which it is applied, thus taking advantage of the translational freedom in setting the origin of the coordinate system.

5.2.2 GAPW Representation of the Induced Current Densities and Gauge

In the GAPW framework, we propose to use for the induced current density a decomposition analogous to the one applied to the electron density,

$$\mathbf{j}(\mathbf{r}) = \tilde{\mathbf{j}}(\mathbf{r}) + \sum_A \Big(\mathbf{j}_A(\mathbf{r}) - \tilde{\mathbf{j}}_A(\mathbf{r})\Big) \tag{5.8}$$

where $\tilde{\mathbf{j}}$ is the soft contribution to the total current density, $\mathbf{j}_A$ is the local hard contribution centered on atom A, and $\tilde{\mathbf{j}}_A$ is the local soft contribution,

which compensates for the double counting.

The convergence of the NMR chemical shifts with respect to Gaussian basis set size is strongly dependent on the choice of the gauge $d(\mathbf{r})$, see e.g. Ref. [18] and references therein. It is thus important to judiciously choose the gauge which provides a good compromise between complexity and convergence property with respect to the basis set size.

The advantage of the $d(\mathbf{r}) = \mathbf{r}$ gauge (CSGT), i.e. that the diamagnetic term is identically zero, can be significantly weakened due to the rich basis set required to obtain an accurate description of the current density close to the nuclei. Thus, it seemed more appropriate to implement the IGAIM approach [19] following the procedure by Cheeseman *et al* [5]. In the IGAIM method, the gauge $d(\mathbf{r})$ is taken as the closest nuclear center.

It is interesting to notice that, in contrast to the chemical shift calculations, the evaluation of the g tensor is less affected by the choice of the gauge even when the simpler and computationally more convenient CSGT approach is used.

5.3 Calculation of the Chemical Shift Tensor

The computation of the chemical shift tensor requires the evaluation of the integral Eq. (5.1). This can efficiently be done using the GAPW representation of the induced current density response.

The contribution to the chemical shift coming from the soft part of the induced current density response $\tilde{\mathbf{j}}$ is computed in reciprocal space. Following the procedure suggested by Pickard and Mauri [37], and Sebastiani and Parrinello [42], we distinguish between the $\mathbf{G} \neq \mathbf{0}$ components and the $\mathbf{G} = \mathbf{0}$ component of the induced magnetic field, where $\mathbf{G}$ denotes a reciprocal space vector. It is observed that the $\mathbf{G} = \mathbf{0}$ component, which cannot be computed within PBC, depends on the macroscopic shape of the studied material. By assuming a spherical shape, it can be approximated computing the magnetic susceptibility arising from the soft induced current density $\tilde{\mathbf{j}}$, as

$$\chi_{xy} = \frac{2\pi}{\Omega_c c} \int \left[\mathbf{r} \times \tilde{\mathbf{j}}_x(\mathbf{r}) \right]_y d^3r \tag{5.9}$$

The contribution to the chemical shift of nucleus A, σ_{xy}^{A}, arising from the induced local current densities is evaluated as

$$\frac{1}{c}\sum_{B}\int_{\Omega_B}\left[\frac{\mathbf{r}-\mathbf{A}}{|\mathbf{r}-\mathbf{A}|^3}\times\left(\mathbf{j}_{xB}(\mathbf{r})-\tilde{\mathbf{j}}_{xB}(\mathbf{r})\right)\right]_y d^3r \tag{5.10}$$

where the sum over B is restricted to nuclei that are within a radius R_c from A. A discussion of the truncation of the summation will be given below. The integration over Ω_B is performed numerically on a spherical grid featuring a logarithmic radial and a Lebedev–type [27, 28, 29] angular discretization. The numerical integration converges rapidly with respect to the number of grid points and about 10000 grid points per atom are enough to converge the chemical shifts below 10^{-1} ppm.

5.4 Results and Discussion

5.4.1 Test Calculations

All developments were implemented in `Quickstep` which is part of `CP2K` [44]. `CP2K` is a freely available (under the GNU General Public License) general program to perform atomistic and molecular simulations of solid state, liquid, molecular and biological systems. A description of `Quickstep` can be found in Ref. [48].

The GAPW method for the AE calculation of the NMR chemical shifts was validated by comparison, for a series of small isolated molecules, with the results from conceptually similar gas–phase methods, namely the NMR routines as implemented in the `g03` program package [11] (further referred to as the IGAIM-GA method).

Isolated molecules are approximated by a supercell approach with a large cell size of $(20\ \text{Å})^3$. We have used a 300 Ry cutoff for the auxiliary plane wave grid, the BLYP [2, 30] gradient–corrected exchange–correlation functional and the efficient and numerically stable orbital transformation energy minimizer introduced in Ref. [50]. The Gaussian basis sets used in this work were taken from the Environmental Molecular Sciences Laboratory (EMSL) basis set exchange library [41].

A comparison between the IGAIM method implemented in `g03` (IGAIM–GA) and the presented IGAIM–GAPW method for the isotropic and the anisotropic chemical shifts of a representative set of small molecules is shown in Fig. (5.1). Three different basis sets were used, namely the Pople split–valence double–zeta plus polarization 6–31G(d,p) [14, 10], and the augmented double and triple–zeta correlation consistent basis sets aug–cc–pVDZ and aug–cc–pVTZ, respectively [9, 52, 21, 51]. The test set is composed of 26 molecules, namely: C_2H_2, CH_2O, CH_3Cl, CH_3F, CH_4, Cl_2, CO_2, F_2, FCl, H_2O_2, H_2O, H_2, H_2S, HCl, HCN, HCOOH, HF, HNO_3, N_2O, N_2, NH_3, NO_2, O_2, O_3, PH_3, and SO_2. All the geometries were optimized at the BLYP/aug–cc–pVDZ level of theory.

For the small 6–31G(d,p) basis set, a maximal unsigned relative error of 30 % is observed for the isotropic chemical shift of the carbon atom in HCN. The isotropic shift for the nitrogen atom in HCN is 3.9 ppm (parts per million) for IGAIM–GA and 5.1 ppm for IGAIM–GAPW. This absolute error (1.2 ppm) is of the same order as for other nitrogen atoms with e.g. 0.6 ppm for NH_3. Except for this extreme case, the other isotropic (anisotropic) chemical shifts have a smaller maximal absolute relative error e.g. less than 1 % (10 %) for hydrogen, 5 % (0.5 %) for carbon and 2 % (2 %) for oxygen.

The relative error is greatly reduced for the larger aug–cc–pVDZ and aug–cc–pVTZ basis sets. The aug–cc–pVDZ provides a maximal absolute relative error of 5 % (nitrogen in HCN) and 10 % (H_2) for, respectively, the isotropic and anisotropic component of the shift. The maximal relative errors for the isotropic shift are 0.5 % for hydrogen, 0.9 % for carbon, 4.8 % for nitrogen, 1.3 % for oxygen, smaller than 0.1 % for fluorine, silicon, and phosphorus, 0.1 % for sulfur, and 0.2 % for chlorine. The maximal relative error for the aug–cc–pVTZ basis set is 0.7 % and 2.5 % for the iso– and anisotropic part of the shift, respectively. For the different atomic kinds, we obtain a maximal relative error of 0.6 % (H), 0.6 % (C), 0.7 % (N), 0.4 % (O), and finally less than 0.1 % (F, P, S and Cl). The corresponding maximal absolute errors are 0.2 ppm (H), 0.1 ppm (C), 0.3 ppm (N), 0.6 ppm (O), 0.3 ppm (F), 0.3 ppm (P), 0.2 ppm (S) and 0.1 ppm (Cl).

The relative accuracy of the chemical shifts mirrors the previous observations for the total energy [17]. This inaccuracy can be traced back to the incompleteness of the basis used to represent the local contribution of the

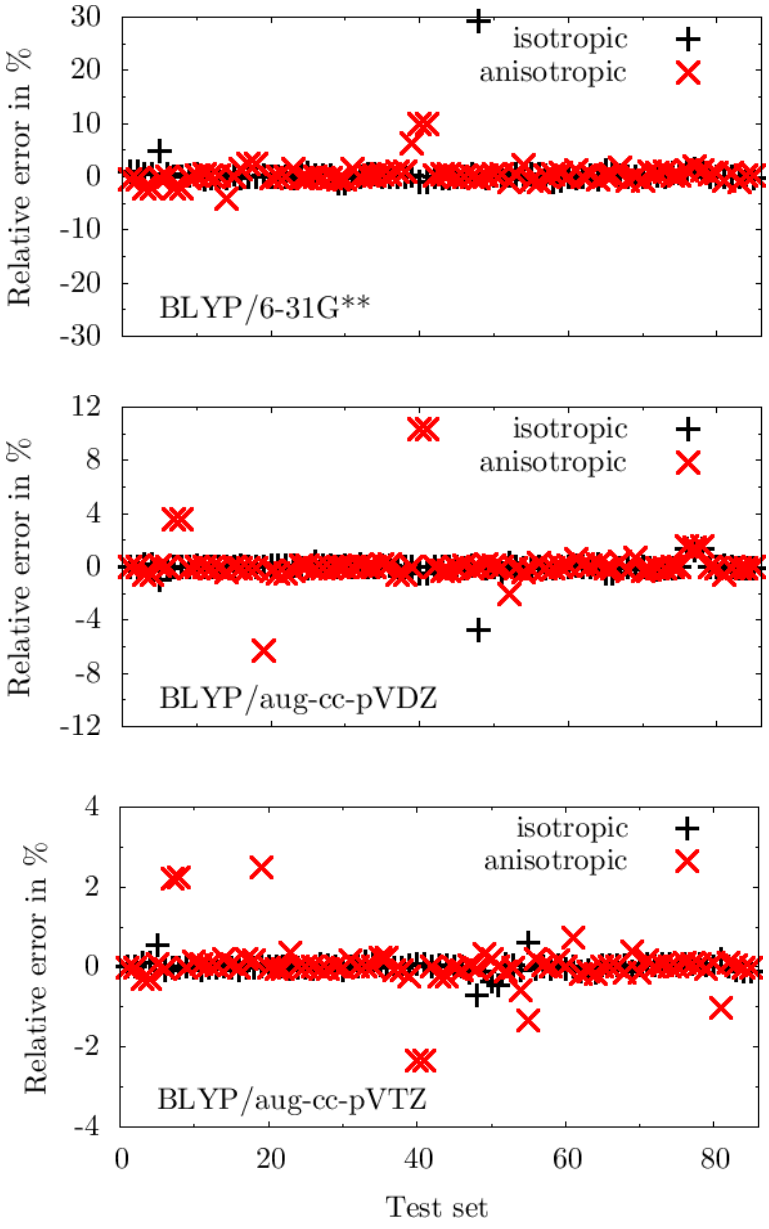


Figure 5.1: Relative error of the IGAIM–GAPW method with respect to the IGAIM–GA method for the isotropic and anisotropic chemical shifts calculated on a set of small molecules. Note the different scales in the three panels.

atomic densities and the induced current densities. In the present implementation, this local basis is constructed from the primitive functions of the original basis set [31, 25] and shows better convergence with increasing the

quality of the basis set. Accordingly, the inaccuracy of the chemical shifts can be systematically reduced by increasing the quality of the basis employed to represent the local contributions.

For a given basis set, the inaccuracy depends on different factors such as the hybridization of the atom of interest and its chemical environment. Therefore, the accuracy of relative shifts might not be systematically improved.

Figure 5.2: Water chain test case for convergence studies.

In Table (5.1), we present the convergence of the isotropic chemical shifts for a test case depicted in Fig. (5.2): the central molecule of a linear water chain with respect to the truncation parameter R_c introduced in Eq. (5.10). The calculations were carried out at the BLYP/aug–cc–pVTZ level of theory. This test system is composed of a linear chain consisting of 9 hydrogen–bonded water molecules (both intramolecular O−H bond lengths of 0.95 Å and H−O−H angle of 109.47 °) with an O−O distance of 3 Å. The O−H_1 bonds are aligned along the z–axis.

Table 5.1: Convergence of the isotropic chemical shifts of the central water molecule of a water chain (see text and Fig. (5.2) for details) with respect to the truncation parameter R_c introduced in Eq. (5.10). The H_1 atoms is aligned along the z–axis. All the values are in ppm.

	$R_c = 0.5$ Å	$R_c = 1.6$ Å	$R_c = 4.0$ Å	all atoms
O	322.2247	322.2465	322.2489	322.2481
H_1	29.8431	29.5129	29.5829	29.5926
H_2	32.8875	32.1919	32.1276	32.1141

Different radial cutoffs are chosen, increasing in size, to include in the summation (Eq. (5.10)), first the single atom where the shift is measured, then the central water molecule, the central and its first neighboring water molecules and finally the full system. While the chemical shifts of the oxygen

are not too sensitive to the truncation, the shifts of the proton are affected by small value of R_c with a worst case difference of up to 0.8 ppm for H_2. A truncation of $R_c = 4$ Å gives error in the chemical shifts for the proton less than 0.02 ppm. Except for the water chain test case, no truncation is used in the rest of this work.

5.4.2 Chemical Shifts of Isolated and Hydrated Adenine

To further validate the method, we present, in Table (5.2), the calculated chemical shifts for an isolated adenine. While the ^{15}N chemical shifts are referenced to $MeNO_2$, the ^{1}H and ^{13}C are given with respect to tetramethylsilane (TMS). The absolute chemical shifts for all the references are also reported in the table 5.2. The geometries of the adenine and all the references were optimized at the BLYP/6–31G(d,p) level of theory. For the labeling scheme of the adenine see figure 5.3.

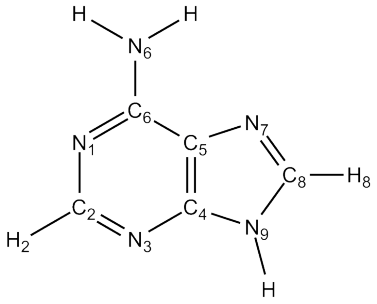


Figure 5.3: Labeling scheme for the adenine.

For comparison, we also report a calculation with the method by Sebastiani and Parrinello (referred to as CSGT–CP) as implemented in `cpmd` [46]. In the calculations the BLYP functional, a plane wave cutoff of 200 Ry and pseudopotential of the Goedecker type [15] were used. Chemical shifts corrected for the core electrons are also presented in parenthesis. The core contribution to the pseudopotential calculations is assumed to be constant for an atom in a chemically "equivalent" environment.

These corrections are calculated as [37]:

$$\delta(X) = \sigma(X_{\text{ref}}) - \sigma(X) + \delta(X_{\text{ref}}),$$

where X is the atom considered, X_{ref} is the *same* atom in a reference system, the $\sigma(X_{\mathrm{ref}})$ and $\sigma(X)$ are computed at the same level of theory and $\delta(X_{\mathrm{ref}})$ is the chemical shift of the reference system with respect to the external reference (here TMS or $MeNO_2$). In this work the $\delta(X_{\mathrm{ref}})$ were calculated at the BLYP/aug–cc–pV5Z level of theory with the IGAIM–GA method. The reference systems are chosen to be benzene for all the carbon atoms, pyridine for N_1, N_3 and N_7, aniline for N_6 and pyrrole for N_9.

Table 5.2: Calculated ^{13}C, ^{1}H and ^{15}N, chemical shifts of an isolated adenine (upper part of the table). Values in parentheses are corrected for core electrons (see text for details). Absolute chemical shifts of the references are shown in the lower part of the table. The subscript t, n, b, py, pr and a refer to TMS, $MeNO_2$, benzene, pyridine, pyrrole and aniline, respectively. All the values are in ppm.

	IGAIM GAPW[1]	IGAIM GA[1]	IGAIM GA[2]	CSGT CP[3]	IGAIM GAPW[4]	GIPAW[5]
C_2	164	164	166	136 (168)	162	166
C_4	159	159	162	130 (161)	156	162
C_5	128	128	130	93 (124)	126	131
C_6	163	163	165	134 (166)	161	164
C_8	141	141	143	115 (146)	140	143
H_2	8.4	8.4	8.5	7.4	8.4	8.5
H_8	7.5	7.5	7.7	7.0	7.6	8.1
N_1	−134	−134	−135	−143 (−118)	−128	−132
N_3	−142	−142	−146	−152 (−127)	−136	−141
N_6	−319	−319	−327	−285 (−332)	−310	−322
N_7	−129	−129	−132	−154 (−129)	−123	−128
N_9	−235	−235	−240	−217 (−240)	−226	−234
C_t	177	177	175	7	181	184
H_t	31.3	31.3	31.3	30.6	31.1	30.9
N_n	−159	−159	−166	−299	−152	−139
C_b			37	−99		
N_{py}			−119	−227		
N_{pr}			77	−79		
N_a			171	−8		

[1] BLYP/cc–pVQZ, [2] BLYP/aug–cc–pV5Z, [3] BLYP/200 Ry, [4] PBE/cc–pVQZ, [5] PBE/100 Ry.

The IGAIM–GA and the IGAIM–GAPW calculations with the cc–pVQZ basis set give similar chemical shifts as expected from the benchmark shown above. While the ^{15}N and ^{13}C chemical shifts calculated with the CSGT–CP method can be in disagreement by up to 40 ppm with respect to the IGAIM–GA/aug–cc–pV5Z calculation, the corrected shifts are in much better agreement (except for N_1 with a difference of about 17 ppm) with the Gaussian–based methods. For the ^{1}H chemical shifts, a large difference (up to 1 ppm) is observed between the Gaussian based and the CSGT–CP methods.

In the last column of table 5.2, we also present the chemical shifts obtained with the GIPAW method, as implemented in the `Quantum-ESPRESSO` program package [47]. The PBE functional, the Troullier–Martins [45] GIPAW pseudopotentials, a $3 \times 3 \times 3$ k–point mesh and a 100 Ry plane wave cutoff were used for the calculations. For the sake of comparison, we also show the chemical shifts obtained with the IGAIM–GAPW method at the PBE/cc–pVQZ level of theory. The shifts obtained with the IGAIM and GIPAW methods are in very good agreement. A deviation of 6 ppm and 12 ppm can be seen for the carbon and nitrogen atoms, respectively. This disagreement is attributed to the quality of the basis set used for the IGAIM–GAPW calculation. The deviation for the chemical shift of the H_8 is surprisingly large (0.5 ppm).

To show further the capability of the implementation of the IGAIM–GAPW method, we present in Table (5.3) the chemical shifts of hydrated adenine calculated within the QM/MM framework and PBC. The full system contains the adenine and 827 water molecules. The coordinates of the system were extracted from a classical molecular dynamics (see Ref. [24] for more details about the setup). The absolute chemical shifts of the reference systems, TMS and $MeNO_2$, were taken from Table (5.2).

The first calculation, reported in the table, consists of the adenine without the solvent (further referred to as ISO) at the BLYP/cc–pVQZ level of theory. The next columns are obtained within the QM/MM framework with different QM regions, *i.e.* the adenine only (W0), and the adenine plus the water molecules overlapping spheres of 3 Å radius (W3) around each solute atoms and the same setup but with a radius of 5 Å (W5). We note that the ^{13}C chemical shifts are mostly insensitive to the presence of the solvent with a maximal variation of about 9 ppm between the ISO and W3 calculations.

Table 5.3: Calculated ^{13}C, ^{1}H and ^{15}N, chemical shifts of adenine at the BLYP/cc–pVQZ level of theory within a QM/MM framework (see text for details). All the values are in ppm.

	ISO	W0	W3	W5
C_2	164	164	166	166
C_4	148	148	148	148
C_5	120	120	120	120
C_6	160	160	159	159
C_8	145	152	154	154
H_2	7.8	7.8	7.8	7.9
H_8	7.7	8.3	8.2	8.4
N_1	−115	−129	−128	−125
N_3	−127	−147	−144	−145
N_6	−330	−330	−318	−317
N_7	−121	−144	−147	−149
N_9	−249	−237	−226	−223

Larger variations between the ISO and W3 calculations are observed for the hydrogen and nitrogen atoms with 0.6 ppm and 26 ppm, respectively. The last column of the table contains the W5 calculations. The difference between the W3 and W5 is very small with less than 1 ppm for the chemical shift of the carbons, 0.2 ppm for the hydrogens and 3 ppm for the nitrogens.

While the maximal variation of the chemical shifts with respect to the size of the QM part is small with about 2 ppm for carbon, 0.1 ppm for hydrogen and 3 ppm for the nitrogen without any N−H bond (N_1, N_3 and N_7), it can be more pronounced with up to 14 ppm for the other nitrogens (N_6 and N_9). This finding reflects the formation of hydrogen bonds between the NH and NH_2 groups with the solvent that are not properly described in the W0 calculation.

5.5 Summary

We have introduced a method for the AE calculation of the NMR chemical shifts with PBC, using the GAPW method. Thanks to the AE–GAPW

scheme, we can avoid the use of the pseudopotential approximation, which is one of the main sources of inaccuracies for the calculations performed with the original Sebastiani and Parrinello implementation, in particular for elements heavier than hydrogen. The method has been validated over a set of small isolated molecules by comparing the deviation of the chemical shifts with other theoretical methods. Then, an applications of the method have been presented, which involves QM/MM calculations of the chemical shifts of an adenine molecule hydrated in a box of 827 water molecules with up to 66 atoms in the QM part. This example is illustrative of the application field in which we hope the proposed method to be of great value.

References

[1] S. Baroni, P. Giannozzi, and A. Testa. Green's–function approach to linear response in solids. *Phys. Rev. Lett.*, 58:1861, 1987.

[2] A.D. Becke. Density–functional exchange–energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098, 1988.

[3] G. Berghold, C.J. Mundy, A.H. Romero, J. Hutter, and M. Parrinello. *Phys. Rev. B*, 61:10040, 2000.

[4] C. Brouder, G. Panati, M. Calandra, C. Mourougane, and N. Marzari. Exponential localization of wannier functions in insulators. *Phys. Rev. Lett.*, 98:46402, 2007.

[5] J.R. Cheeseman, G.W. Trucks, T.A. Keith, and M.J. Frisch. *J. Chem. Phys.*, 104:5497, 1996.

[6] R. Declerck, E. Pauwels, V. Van Speybroeck, and M. Waroquier. First–principles calculations of hyperfine parameters with the Gaussian and augmented–plane–wave method: Application to radicals embedded in a crystalline environment. *Phys. Rev. B*, 74:245103, 2006.

[7] R. Declerck, V. Van Speybroeck, and M. Waroquier. First–principles calculation of the EPR g tensor in extended periodic systems. *Phys. Rev. B*, 73:115113, 2006.

[8] R. Ditchfield. Self–consistent perturbation theory of diamagnetism. *Mol. Phys.*, 27:789, 1974.

[9] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90:1007, 1989.

[10] M.M. Francl, W.J. Petro, W.J. Hehre, J.S. Binkley, M.S. Gordon, D.J. DeFrees, and J.A. Pople. Self–consistent molecular orbital methods. xxiii. a polarization–type basis set for second–row elements. *J. Chem. Phys.*, 77:3654, 1982.

[11] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin,

J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[12] X. Gonze. Adiabatic density–functional perturbation theory. *Phys. Rev. A*, 52:1096, 1995.

[13] X. Gonze. Perturbation expansion of variational principles at arbitrary order. *Phys. Rev. A*, 52:1086, 1995.

[14] P.C. Hariharan and J.A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theo. Chim. Acta*, 28:213, 1973.

[15] C. Hartwigsen, S. Goedecker, and J. Hutter. Relativistic separable dual–space Gaussian pseudopotentials from H to Rn. *Phys. Rev. B*, 58:3641, 1998.

[16] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864, 1964.

[17] M. Iannuzzi, T. Chassaing, T. Wallman, and J. Hutter. Ground and excited state density functional calculations with Gaussian and augmented plane waves method. *Chimia*, 59:499, 2005.

[18] M. Kaupp, M. Bühl, and V. G. Malkin. *Calculations of NMR and EPR parameters: Theory and Applications*. Wiley–VCH, 2004.

[19] T.A. Keith and R.F.W. Bader. Calculation of magnetic response properties using atoms in molecules. *Chem. Phys. Lett.*, 194:1, 1992.

[20] T.A. Keith and R.F.W. Bader. Calculation of magnetic response properties using a continueous set of gauge transformations. *Chem. Phys. Lett.*, 210:223, 1993.

[21] R. A. Kendall, T. H. Dunning, and R. J. Harrison. Electron affinities of the first–row atoms revisited. systematic basis sets and wave functions. *J. Chem. Phys.*, 96:6796, 1992.

[22] W. Koch and M.C. Holthausen. *A chemist's guide to density functional theory.* Wiley–VCH, 2000.

[23] W. Kohn and L.J. Sham. Self–consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133, 1965.

[24] S. Komin, C. Gossens, I. Tavernelli, U. Rothlisberger, and D. Sebastiani. NMR solvent shifts of adenine in aqueous solution from hybrid QM/MM molecular dynamics simulations. *J. Phys. Chem. B*, 111:5225, 2007.

[25] M. Krack and M. Parrinello. All–electron *ab–initio* molecular dynamics. *Phys. Chem. Chem. Phys.*, 2:2105, 2000.

[26] J. Kussmann and C. Ochsenfeld. Linear–scaling method for calculating nuclear magnetic resonance chemical shifts using gauge–including atomic orbitals within Hartree–Fock and density–functional theory. *J. Chem. Phys.*, 127:54103, 2007.

[27] V. I. Lebedev. Values of the nodes and weights of ninth to seventeenth order gauss–markov quadrature formulae invariant under the octahedron group with inversion. *Zh. Vychisl. Mat. Mat. Fiz.*, 15:44, 1975.

[28] V. I. Lebedev. Quadratures on a sphere. *Zh. Vychisl. Mat. Mat. Fiz.*, 10:293, 1976.

[29] V. I. Lebedev. Spherical quadrature formulas exact to orders 25–29. *Sibirsk. Mat. Zh.*, 17:99, 1977.

[30] C. Lee, W. Yang, and R.G. Parr. Development of the Colle–Salvetti correlation–energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785, 1988.

[31] G Lippert, J. Hutter, and M. Parrinello. The Gaussian and augmented–plane–wave density functional method for *ab initio* molecular dynamics simulations. *Theor. Chem. Acc.*, 103:124, 1999.

[32] F. London. Quantum theory of interatomic currents in aromatic compounds. *J. Phys. Radium*, 8:397, 1937.

[33] N. Marzari and D. Vanderbilt. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B*, 56:12847, 1997.

[34] F. Mauri, B. G. Pfrommer, and S. G. Louie. Ab initio theory of NMR chemical shifts in solids and liquids. *Phys. Rev. Lett.*, 77:5300, 1996.

[35] C. Ochsenfeld, J. Kussmann, and F. Koziol. Ab initio NMR spectra for molecular systems with a thousand and more atoms: a linear–scaling method. *Angew. Chem. Int. Ed.*, 43:4485, 2004.

[36] S. Piana, D. Sebastiani, P. Carloni, and M. Parrinello. Ab initio molecular dynamics–based assignment of the protonation state of pepstatin A/HIV–1 protease cleavage site. *J. Am. Chem. Soc.*, 123:8730, 2001.

[37] C.J. Pickard and F. Mauri. All–electron magnetic response with pseudopotentials: Nmr chemical shifts. *Phys. Rev. B*, 63:245101, 2001.

[38] C.J. Pickard and F. Mauri. First–principles theory of the EPR g tensor in solids: defects in quartz. *Phys. Rev. Lett.*, 88:86403, 2002.

[39] A. Putrino, D. Sebastiani, and M. Parrinello. Generalized variational density functional perturbation theory. *J. Chem. Phys.*, 113:7102, 2000.

[40] R. Resta. Quantum–mechanical position operator in extended systems. *Phys. Rev. Lett.*, 80:1800, 1998.

[41] K.L. Schuchardt, B.T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T.L. Windus. Basis set exchange: a community database for computational sciences. *J. Chem. Inf. Model.*, 47:1045, 2007.

[42] D. Sebastiani and M. Parrinello. A new ab–initio approach for NMR chemical shifts in periodic systems. *J. Phys. Chem. A*, 105:1951, 2001.

[43] D. Sebastiani and U. Rothlisberger. Nuclear magnetic resonance (NMR) chemical shifts from hybrid DFT QM/MM calculations. *J. Phys. Chem. B*, 108:2807, 2004.

[44] The CP2K developers group, 2008.

[45] N. Troullier and J. L. Martins. Efficient pseudopotentials for plane–wave calculations. *Phys. Rev. B*, 43:1993, 1991.

[46] `CPMD, Version 3.9`. copyright IBM Corp. 1990–2004, copyright MPI für Festkörperforschung Stuttgart 1997-2001; http://www.cpmd.org/.

[47] `Quantum-ESPRESSO` is a community project for high–quality quantum–simulation software, based on density–functional theory, and coordinated by P. Giannozzi, 2008.

[48] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter. Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comp. Phys. Comm.*, 167:103, 2005.

[49] G.H. Wannier. The structure of electronic excitation levels in insulating crystals. *Phys. Rev.*, 52:191, 1937.

[50] V. Weber, J. VandeVondele, J. Hutter, and A.M.N. Niklasson. Direct energy functional minimization under orthogonality constraints. *J. Chem. Phys.*, 128:84113, 2008.

[51] D.E. Woon and T.H. Dunning. Gaussian basis sets for use in correlated molecular calculations. iii. the atoms aluminum through argon. *J. Chem. Phys.*, 98:1358–1371, 1993.

[52] D.E. Woon and T.H. Dunning. Gaussian basis sets for use in correlated molecular calculations. iv. calculation of static electrical response properties. *J. Chem. Phys.*, 100:2975, 1994.

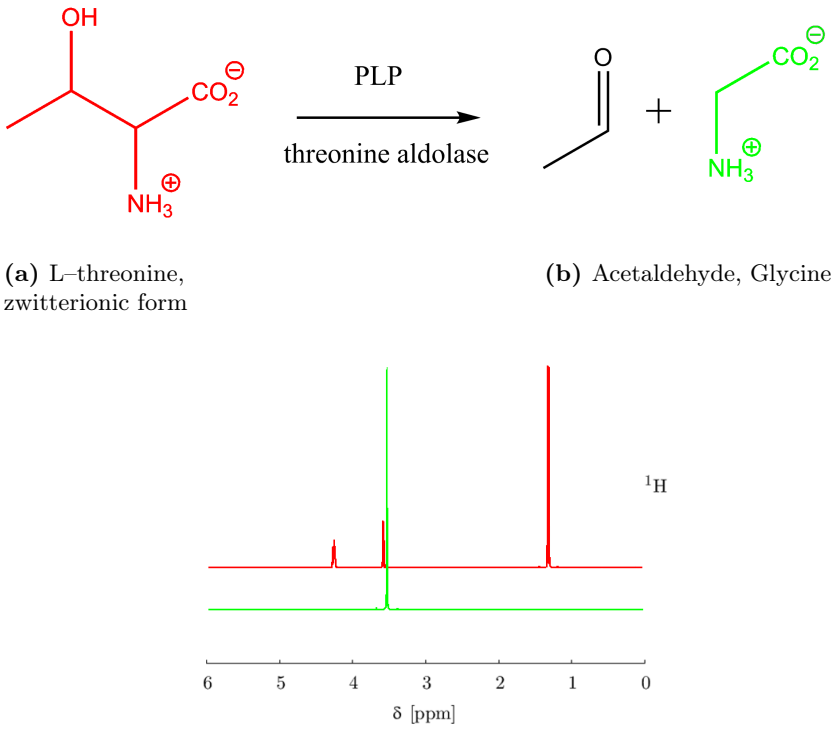
Chapter 6

Aqueous Quantum–Chemical Simulations of NMR Spectra: Amino Acids and Peptides in Their Natural Environment

6.1 Introduction

In a living organism, life is represented by a continuous interplay of chemical transformations. Potentially every bioactive compound could trigger a reaction cascade (i.e. steps in signal transduction), hence leading to a multitude of new chemical substances. Almost every step in a biochemical modification is mediated by the natural catalysts called enzymes. As an example, let us consider one metabolic pathway of the amino acid threonine: in a reaction it finds itself transformed in glycine (Fig. (6.1)). Threonine aldolase catalyzes this reaction via pyridoxal–phosphate (PLP), which is a bioactive form of vitamin B_6.

The simplest enzymatic mechanism can be summarized as follows: the active compound (natural or drug) first docks to the active site of the protein via complementarity principle in enzyme–substrate specific shapes and interactions, and then the catalysis takes place. It is important to stress that the protein–ligand binding process follows the concepts of "lock and



(a) L–threonine, zwitterionic form

(b) Acetaldehyde, Glycine

(c) Threonine (red) and Glycine (green) NMR spectra

Figure 6.1: Upper panel: threonine aldolase mechanism in aqueous phase. This enzyme belongs to the family of lyases, specifically the aldehyde–lyases, which cleave carbon–carbon bonds. It employs one cofactor, pyridoxal–phosphate (PLP). Lower panel: NMR spectrum for threonine and glycine.

key" and "induced fit" mechanisms. The latter principle in particular indicates that enzymes modulates their activity via a certain degree of plasticity, undergoing sensible conformational change during the catalytic process.

Thus, when confronted with the investigation–characterization challenge in biochemistry two tasks are faced: the identification of possibly every single intermediate of a given reaction, the other understanding the nature of the conformational changes that an enzyme undergoes while performing the catalysis.

Spectroscopy is amongst the possible methods for the characterization of substances. Notably, Nuclear Magnetic Resonance (NMR) spectroscopy plays a key role in the context of structure elucidation, in particular for

proteins, both in solution as well as embedded in complex matrix such as membranes. The differentiation power of this method relies on the fact that chemically equivalent groups display distinct signals depending on the environment where they are located.

NMR methods are unique amongst today's available techniques for structure determination of molecules at atomic resolution. This is particularly true if one consider NMR data for compounds present in living organisms, such as proteins and nucleic acids, because of the possibility of recording, like in Fig. (6.1c), spectra in solution [41]. Considering that protein–containing fluids (such as blood, stomach liquid or saliva) are the natural environment where these biomolecules play their physiological roles, a knowledge of the structure adopted in solution is highly desirable. In NMR experiments, solution conditions like temperature, pH and salt concentration can be adjusted so as to closely mimic a given physiological environment. Reversely, the experimental settings may also be changed giving rise to non–physiological conditions, to study for example the process of protein denaturation. The range of information NMR can access is very broad: structure determination, ability to investigate dynamical features experienced by a given molecular structure, studies of thermodynamic and kinetic details of interactions between compounds and other solution components (either macromolecules or low molecular weight ligands). Again, those data can be measured directly in solution, imitating natural milieu.

Chemical shifts data are the most accessible and easily measured quantities in NMR spectroscopy. For proteins in particular, these values reveal detailed information about relevant geometrical parameters [33]: for instance, the correct interpretation of the signals can give indications concerning the backbone dihedral angles ψ and ϕ, sidechain χ angles, hydrogen bond interactions, local electric fields, proximity and orientation of aromatic rings. Furthermore, it allows to discriminate amongst ionization or oxidation states and it can provide even description of internuclear distances. Protein chemical shifts are a source of information in analysis including secondary structure mapping, generating structural constraints, three–dimensional structure refinement and three–dimensional structure generation [30, 6]. This has become possible thanks to methods developed to attribute sequential assignments for mapping a specific resonance transition frequency to a specific atom in the molecule. Examples are site–directed mutagenesis, spin decou-

pling, proton exchange, homo– and hetero–nuclear multidimensional NMR techniques and nuclear Overhauser enhancement [40, 11, 19].

In addition to these features, accurate analysis of the recorded signals give access to the entire picture of dynamical aspects ranging from conformational change, backbone dynamics and phenomena such as ring–flip rates. The temporal evolution of an NMR chemical shift spectrum gives the possibility to follow the kinetics of an enzymatic activity manifested either via ligand binding or via chemical reaction.

However, often the results are difficult to interpret because the intrinsically complexity of the experiments. In this respect computer simulations aims to help to understand existing data and propose new experiments. With numerical calculations possibly effects can be accurately modeled and influence of individual factors can be identified and rationalized.

In this work NMR spectroscopical properties of several water solvated amino acids and peptides are investigated via computer simulations. The crucial aspects that have to be addressed and incorporated in the modeling part, are an atomistic description of the molecular system, where the aqueous phase is appropriately taken into account, and an explicit complete representation of the electronic structure of the system. This in order to mimic experimental conditions allowing a direct comparison with experimental results.

6.2 Computational Details

We are aiming to calculate chemical shift parameters by the means of efficient quantum–chemical calculations based on a recent implementation [37] within an all–electron scheme employing periodic boundary conditions (PBC), in order to asses the accuracy of the methods in the context of solvated biomolecules. This approach is ideal to investigate the contributions that arise from finite temperature effects: namely the continuously changing solvation pattern around a given solute, as well as the multitude of configurations that arise from vibrational effects, making the system under analysis to explore a large configurational space. The strategy is to rely on combinations of molecular dynamics techniques, in order to sample conformational space,

and quantum mechanical calculations to collect spectra along the trajectory. With this computational–affordable framework the time–evolution of the structure is probed and reliable averages can be calculated.

There are few other methods that allow to calculate *ab initio* chemical shifts proposed in the literature devised to deal with large systems [21, 36, 28, 29]. We would like to overcome some of the drawbacks of PBC–based approaches, the most important being to lift the pseudopotential approximation or the *a posteriori* reconstruction of the core electronic density. It is indeed in this respect that, although relying on periodic boundary conditions, our implementation differs from the other existing ones, namely allowing an efficient explicit calculation of the core electron density (see Chapter 5).

Due to molecular motion, liquid systems undergo substantial reorganization and the solute is experiencing a large variety of solvation patterns. In the case of an aqueous solution these solvation patterns are linked to the aptitude of interaction via hydrogen bonds with the surrounding waters. These H–bonds can further extend far from the solute molecule, involving chain of waters beyond the simple first neighboring solvent molecules. Furthermore, this first sphere of solvation can undergo substantial reorganization *per se* as well as exchanging with the second sphere, modifying transiently the amount of water molecules coordinating the solute.

Because NMR experiments are correlated with spin relaxations process, intrinsically the timescale is bounded to typically micro to milliseconds. This limit is dictated by the magnitude of relaxation times and leads to the fact that what is observed in NMR spectra are averages coming from multiple structures. This is undoubtedly revealed if one focuses on exchange mechanisms, in particularly beyond the point of coalescence, where signals collapse due to the fast kinetic of chemical exchange. It is thus essential to take into account motional effects in the calculation [20, 9].

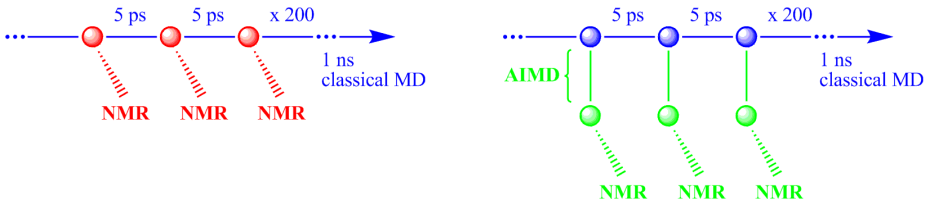
General protocols for calculating NMR chemical shifts based on sampling along a trajectory generated by molecular dynamics (MD) algorithms are gaining more and more consensus, as they allow the generation of a distribution of geometries according to a given thermodynamical ensemble [9, 4]. The chemical shift can be finally calculated on top of selected configurations. There are many computational models attempting to address the microscopic dynamical behavior of systems. They are based on classical molecular me-

chanics (MM), *ab initio* (QM) and combination of thereof, giving rise to hybrid classical–quantum mechanic (QM/MM) methods. Bearing in mind this distinction, depending which type of calculations are performed, this dictates the affordable size and the timescales accessible of the simulation. This aspect is also reflected in the quality of the potential employed to describe the interactions between the particles. On the MM side, we rely on a force fields for biomolecules in aqueous phase called Amber [5], in particular `amber.ff03` and TIP3P water, with 1 fs timestep for the evolution of the equation of motion at 300 K. On the QM side, the level of theory adopted here is density functional theory (DFT), which has been proved to provide accurate MD properties for solvated systems [16]. For the MD simulations we use the QM/MM approach with the Becke–Lee–Yang–Parr (BLYP) exchange–correlation functional [1, 22], and the TZV2P–GTH basis set. The plane wave cutoff used is 300 Ry, which is practically at convergence for a GPW calculation. 0.5 fs timestep for propagating the equations of motion at 300 K in the NVE ensemble. The NMR chemical shift calculations are performed in the framework of DFT and linear response theory, where the Dunning correlation–consistent quadruple–zeta basis set cc–pVQZ, 300 Ry and BLYP are used [24]. The `CP2K` software package is an open source implementation of the aforementioned approaches, and was used for molecular dynamics, energetic and spectral calculations presented in this work.

As pointed out, typically, a NMR experiment takes place in a timescale between the micro– and the millisecond. In the context of mimicking physiological conditions, experimentalists operate in aqueous solution at 300 K. Just like in a real experiment, where a given measurement is realized many times, we are aiming to perform statistical analysis on the time evolution of a solute structure–dependent NMR signal along a MD trajectory according to an equilibrium thermodynamical ensemble. Conventional trajectories can be classical or *ab initio* generated. Besides these two approaches, we consider here also a variation of *ab initio*–derived MD (AIMD) based on snapshots of classical MD as a starting point, with the aim of locally and consistently ameliorate the description of interatomic distances and capture fast oscillations effects.

The sampling sequence for both methods is schematically represented in Fig. (6.2). At equidistant steps along a classical MD trajectory, either snapshots of the simulation are directly taken for NMR calculations (Fig. (6.2a);

or taken as a basis for a short AIMD propagation and subsequently for NMR calculations (Fig. (6.2b)).



(a) classical MD **(b)** *ab initio*–derived (AIMD)

Figure 6.2: Sampling protocols: (6.2a) sampling directly from classical trajectory. (6.2b) sampling from QM/MM trajectories derived from the original classical trajectory.

6.3 Simulation of Amino Acids Spectra

The common structural leitmotif of the amino acid class of molecules is that they possess an amine and a carboxyl functional group, both bonded to the same carbon atom: the alpha carbon (C^{α}). In living organisms, this latter is also bonded to an hydrogen and to a so–called lateral chain, which, differing in composition, gives rise to the 18 natural occurring standard amino acids.

One of the simplest classification of the amino acids (AA) follows the nature of the lateral chain. It consists in regrouping into categories AA molecules that share similarities and give rise to similar properties. These families of lateral chains are: aliphatic, cyclic, hydroxyl– or sulfur–containing side chains, basic or acid, aromatic AA.

Focusing on a plausible structure of an AA in solution, there are mainly two sources of debate. The first is related to the structural arrangement of the lateral residue. Indeed, the lateral chain, can be ramified and can go up to 7 heavy–atoms, which makes this class of molecules substantially floppy, whose conformation may span over large region of phase space. The second consideration is related to the protonation state, dealing with acidic and basic groups within the molecules. This issue is usually solved thanks to experimental pKa's values of a given moiety within the AA, allowing to

systematically rule out protonation states which are not compatibles with a given pH condition. We focus on not charged residues and in zwitterionic form at neutral pH.

6.3.1 Simple Test Cases: Solvated Glycine and Ethanol

Table 6.1: Basis set dependence: absolute ^{1}H and ^{13}C chemical shift of snapshot nr. 1 (first frame) of glycine for the basis set cc–pV(X)Z, with X=D,T,Q and 5. Data columns denoted (aq) indicate the presence of explicit water solvation, whereas data denoted (g) refers to the same glycine geometry where the waters have been stripped off. Results are in ppm.

	cc–pVDZ		cc–pVTZ		cc–pVQZ		cc–pV5Z	
	(g)	(aq)	(g)	(aq)	(g)	(aq)	(g)	(aq)
C^{α}	143.13	158.54	135.04	135.00	132.27	129.87	130.81	129.74
C	23.94	49.54	18.35	20.61	10.64	6.88	6.60	3.15
$H^{\alpha 2}$	28.01	30.80	28.05	30.04	27.87	29.44	27.80	29.08
$H^{\alpha 3}$	27.82	36.05	28.05	28.72	27.99	27.72	27.97	27.90

A first system used for benchmarking purposes is glycine solvated by water. Glycine is the simplest amino acid and its chemical structure in solution is a zwitterion. The QM region integrates 10 atoms, while there are 2800 water molecules in the MM part. Basis set (BS) dependence of the absolute chemical shift, as illustrated in Table (6.1), summarizes that, for the same glycine conformation, two situations can be distinguished, namely with or without water solvation. For the condensed phase, the passage from double–zeta to triple–zeta BS results in a $\sim$ 25 ppm variation for C atoms and up to several units of ppm for H atomic specie. By increasing the expansion from to triple–zeta to quadruple–zeta these variations sets in the regime of few ppm for carbons and a fraction of a ppm for hydrogens, whilst for the isolated phase this levels of magnitude are already reached from the first series of basis set expansion. By further increase of the quality of the BS up to quintuple–zeta the variations with respect quadruple–zeta results in a more subtle tuning of the values: this is still in the order of a few units of ppm for C and a tenth of ppm for H for both phases, indicating systematically that

with cc–pVQZ this level of accuracy is already reached. This findings are consistent with previous studies of Case et al. on extrapolation to complete basis set limit [27].

Table (6.2) summarizes the relative chemical shift of glycine to a reference compound, tetramethylsilane, abbreviated TMS (g) (see later in text for a detailed explanation). The convergence of the relative chemical shift with respect to the BS is somehow faster compared to the absolute one (Table 6.1). Already by changing to triple–zeta to quadruple–zeta the magnitude of the relative change is in the order of few ppm for C and a fraction of ppm for H. This is due to error compensation.

Table 6.2: Basis set dependence: TMS(g) relative ^{1}H and ^{13}C chemical shift of snapshot nr. 1 (first frame) of glycine for the basis set cc–pV(X)Z, with X=D,T,Q and 5. Data columns denoted (aq) indicate the presence of explicit water solvation, whereas data denoted (g) refers to the same glycine geometry where the waters have been stripped off. Results are in ppm.

	cc–pVTZ		cc–pVQZ		cc–pV5Z	
	(g)	(aq)	(g)	(aq)	(g)	(aq)
C^{α}	46.59	46.63	46.60	49.00	46.34	47.41
C	163.28	161.02	168.23	171.99	170.55	174.00
$H^{\alpha 2}$	3.47	1.48	3.74	2.17	3.77	2.49
$H^{\alpha 3}$	3.47	2.80	3.62	3.89	3.60	3.67

In other words, a cc–pV5Z expansion would lead to a modification compared to cc–pVQZ of a few percent for both atomic species. We decide to adopt quadruple–zeta basis set expansion for all further calculations.

For predictive applications, a chemical shift accuracy of approximately $\pm$ 0.1 ppm and $\pm$ 1 ppm for H and C, respectively, are necessary [31].

The effect of the level of theory of the calculations can be rationalized via the analysis of Table (6.3), which illustrates the absolute chemical shift of a selected geometry of glycine (the same snapshot of Table (6.1) with water coordinates stripped off). Several software packages have been used to perform the calculations, `CP2K`, `g03` program package [12] allowing DFT

Table 6.3: Theory dependence: absolute ^{1}H and ^{13}C chemical shift of snapshot nr. 1 of glycine (lacking of explicit water solvation), according to different levels of theoretical description, all using the cc–pVQZ basis set. Results are in ppm.

	CP2K[1]	g03[1]		CFOUR[2]	
	BLYP	BLYP	B3LYP	CCSD	CCSD(T)
C^{α}	132.27	132.47	133.74	148.54	148.40
C	10.64	10.91	9.70	22.37	21.82
$H^{\alpha 2}$	27.87	27.87	27.88	28.09	28.11
$H^{\alpha 3}$	27.99	27.94	27.92	28.10	28.10

Gauge used for the NMR shift calculations: [1] IGAIM, [2] GIAO.

hybrid functional NMR calculations and CFOUR [32] where high–level quantum chemical coupled–cluster techniques are implemented (i.e. CCSD and CCSD(T) approximations).

The divergence between cc–pVQZ/BLYP results obtained from CP2K and g03 are negligible in essence, with a discrepancy between them of 0.2 ppm for carbon and less than 0.05 ppm for hydrogen. g03 BLYP *versus* g03 B3LYP difference amounts to 1.2 ppm for both carbons and 0.01/0.02 ppm for hydrogens. Although CP2K and g03 rely on IGAIM method for the gauge choice and CFOUR on GIAO, a comparison between the two kinds of methods is validated by examining Hartree–Fock obtained NMR values for both g03 and CFOUR, indicating that a convergence towards the same values is reached despite the different choice of the gauge. From a theoretical point of view, different gauge choices should lead to the same value: this is true, however, in the limit of a complete basis set. The difference between the two HF implementations being 1.5 ppm and 0.01 ppm for carbons and hydrogens respectively (data not shown). CCSD and CCSD(T) results are almost identical, with the margins of fractions of ppm. CP2K and CCSD or CCSD(T) values for C and H differ by 16–11 ppm and 0.2 ppm.

As a general consideration, another factor that could affect the NMR value of a given atomic site in QM/MM calculations is related to the extension of the QM region. The solvated data discussed previously in Table (6.1) are

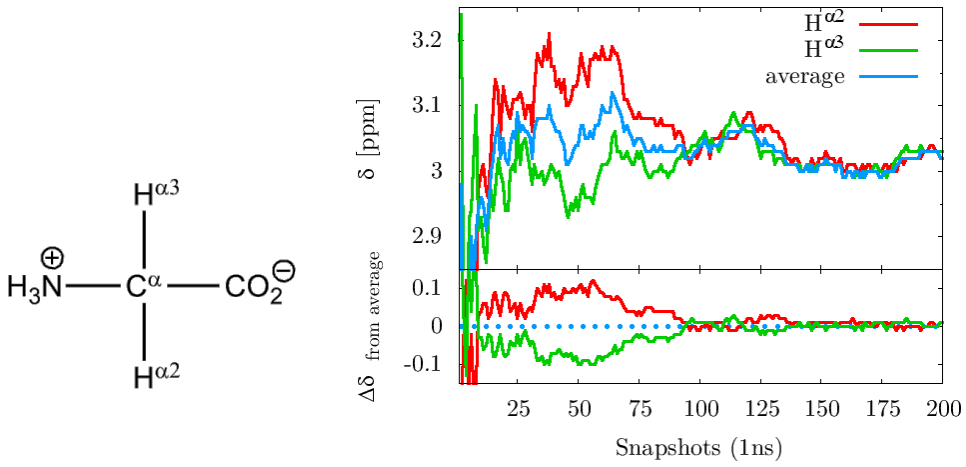
related to a solute–only QM topology, where only glycine were described quantum mechanically whereas water solvent classically. This system is denoted as W0. Analyzing a given configuration (the same snapshot nr.1 used for the previous benchmarking), the QM region can be increased, incorporating progressively waters within a radius of 3 Å (called W3) or 5 Å (called W5) of distance from the solute. 15 respectively 45 water molecules are added to the quantum description. The results indicates that a small difference is found between extended QM systems W3, W5 and W0, accounting for a deviation between all the three systems is in the order of 0.2 ppm for H and less than 1 ppm for C.

By running a molecular dynamics (MD) trajectory, the aim is to sample configurations or micro–states that are representative of a given thermodynamical ensemble. To assess and validate this assumption other factors need to be considered. Firstly, special care have to be devoted to the so–called "equilibration period" before being able to actually go in the production run and collect the data. Secondarily, a sort of "proof of principles" condition related to molecular symmetry has to be fulfilled: it is easy to recognized that there are atoms thought to be NMR identical and, on this respect, particular checks have to be carried out on the chemical environment of those candidates. Supposing that some atomic sites are chemically identical, it translates into the fact that they must experience similar intra– and intermolecular environments. In the structure of aqueous glycine, as depicted in Fig. (6.3a), there are several equivalent atoms. The three amino hydrogens, the two oxygens and the two α hydrogens.

A simple test on the oxygens, from the geometrical point of view, can be to monitor the N–C^{α}–C–O dihedral angle, checking if during the time window of the simulation a rotation around this bond is observed, scrambling the position of the two equivalent terminal oxygens. As an intermolecular descriptor the hydrogen bond pattern can be taken. This parameter is directly related to solvation effects, and for equivalent sites the water shells experienced should be very similar (criterion for first shell < 3.4 Å, second shell < 5 Å of atomic distance of a D–H$\cdots$A hydrogen–bond).

By performing the aforementioned tests for the simple cases such as glycine and ethanol we have indications that classical MD gives sufficient site exchange and the same solvation is probed. This formal verifications are re-

flected into the fact that the average value of the chemical shift are the same for the considered atoms, testifying that this kind of approach allows to describe correctly equivalent atomic species. This both within the timeframe of the simulation, the combination of used parameters as well as the solvent relaxation time occurring after a given event like a bond rotation.



(a) **(b)**

Figure 6.3: Glycine(aq): (6.3a) zwitterionic structure indicating the labeling scheme. (6.3b) H^{α}'s convergence: floating average of $H^{\alpha 2}$, $H^{\alpha 3}$ and their average during an MD trajectory (upper panel), difference between the average and their individual H^{α}'s floating average (lower panel).

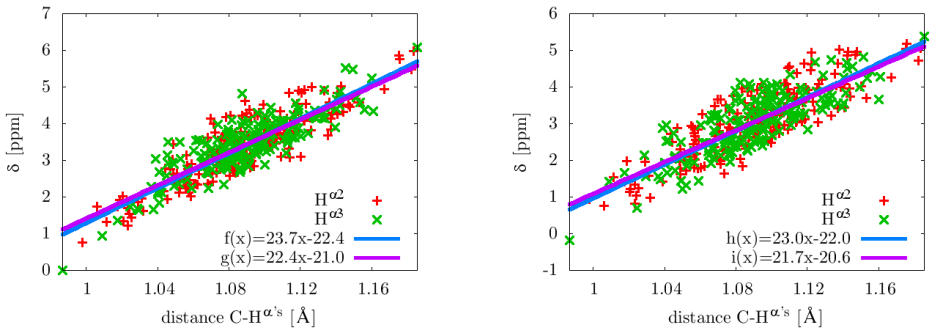
The purpose of Fig. (6.3b) is to illustrate the convergence behavior of the calculated NMR signal on two equivalent atoms, i.e. $H^{\alpha 2}$ and $H^{\alpha 3}$ of glycine. Data are collected along a 1 ns of a classical MD trajectory by sampling configurations every 5 ps. This separation could be considered appropriate considering the rotational correlation time for water approximately of 2 ps. The instantaneous value of the chemical shift undergoes an excursion around the average value. A criterion to assess the convergence is to consider the value of the average while adding more and more points to the results distribution. The floating (or cumulative) average along the sampling represents sort to say the "real time" average while the snapshots are generated [4]. This allow to identify, if any, a plateau region which indicates that convergence is reached. Applying this criterion is related to the principle that by incorporating more and more data in the calculation of the

average, this latter finds itself unchanged. For a two sites example like the alpha hydrogens in glycine, is observed that the "total" average, namely the average of the two converged values, coincides to the overall average of the individual H^{α}'s.

Results obtained from averaging from MD simulation can be poised by a time correlation, that can be identified with a block averaging [17]. The method involves repeated "blocking" of data, and computation of increasing lower bounds for the standard deviation [10]. Such analysis shows that adopting a sampling distance of 5 ps leads to uncorrelated data.

Fig. (6.3b) illustrates the chemical shielding relative to TMS(g) (see further explanation in text), bringing the absolute chemical shift down to the usual ^{1}H range between 0 and 10–12 ppm.

Focusing on the 200 snapshots calculated glycine proton spectra, attempting to correlate the results with some simple geometrical parameter, Fig. (6.4) depicts the value of the chemical shift (also denoted δ) for $H^{\alpha 2}$ and $H^{\alpha 3}$ of a given snapshot and their distance to the C^{α}, the atom to which they are directly bounded to.



(a) **(b)**

Figure 6.4: Distance and chemical shift correlation: (6.4b) indicates solvated glycine, while (6.4a) to not–solvated phase. Both graphs refers to the same trajectory where in the second case the water molecules are simply stripped off. Solute geometries are thus identical for the two calculations.

Both graphs in Figure (6.4) display naïvely that a certain relation is present between hydrogen–heavy atom distance and the NMR signal. A linear interpolation procedure has been used to fit individual H^{α} data. The

functions f(x) and g(x) in Fig. (6.4a) refers to the atoms in naked glycine, whereas h(x) and i(x) in Fig. (6.4b) to the solvated molecule.

A comparison between the case of solvated glycine, Fig. (6.4b), and the case of the same glycine conformation where the waters have been deleted Fig. (6.4a) indicates that, for an identical structure conformation, the presence of the explicit solvation induce a spread of the distribution of the results. In fact, by going from naked glycine to solvated phase, the correlation coefficient of the linear regression is slightly lowered, indicating a better fit for the first case of figure (lines f(x) and g(x) rather than h(x) and i(x)). Furthermore the effect of solvation gives chemical shifts that are 0.5 ppm lower compared to glycine–only.

What the two calculations have in common, although the fundamental difference regarding the explicit water solvation, is that by increasing the interatomic distance, the corresponding chemical shift δ value also increases. As expected, the larger the interatomic distance, the higher the polarization induced to the electronic density generating a deshielding phenomenon. This clear trend is witnessed by the positive slope of the interpolation functions. The magnitude of those slopes is around 22.5 ppm per Å, which translated into 0.225 ppm per 0.01 Å, indicating that a small deviation of the interatomic distance of the conformation used for the calculation do exerts a significant impact of the outcome of the NMR simulated value. The excursion of the C–H^{α}'s distance value of this MD trajectory is $\sim$ 0.15 Å, having a fluctuation of the chemical shift that spans 5 ppm for every H^{α}.

The repercussion of this finding is that, possibly, the main factor affecting NMR calculations can be related to the inaccuracy of the generate structures coming from the sampling.

To test and verify this hypothesis we investigate the contribution to the NMR chemical shift arising from adopting different sampling techniques.

With the purpose to examine the impact exerted by different parametrization governing the interaction potential between the solute atoms, the effects on the glycine MD of another force field [13] are investigated. The two force fields differs by the partial charges assigned to the atomic kinds. The outcome of the chemical shift results arising from structures generated from the two classical potentials (Amber is indicated by ff #1 while the second

by ff #2) is depicted in Fig. (6.5a). Essentially they converge to the same value.

Another setup is to adopt a QM/MM combination and carry out *ab initio* MD (AIMD) from the snapshots of classical MD. This latter approach has been detailed in the introductory section (see Fig. (6.2b)). We recall here that basically on top of a classical MD trajectory, several independent AIMD are propagated for 100 fs. After this period the final configuration is than taken as a basis for NMR calculation. These AIMD are derived from the same original MD but there is no continuity between two successive AIMD trajectories, because their starting configuration is 5 ps apart on the MD potential energy surface.

Both MD and AIMD results are evaluated and presented in the forthcoming tables.

Another strategy that could in principle be adopted is to perform a geometry optimization of every used MD structure and afterward calculate the chemical shift, aiming to restore solute equilibrium geometry. The disadvantage of this procedure is that by optimization the structures would evolve towards restoring all the H–bonds that under the effects of the temperature have been transiently disrupted. This procedure has thus been discarded.

Since absolute shifts are rarely needed, it is common practice to define the chemical shift in terms of the difference in resonance frequencies between the nucleus of interest ν and a reference nucleus ν_{ref}, by means of a dimensionless parameter δ:

$$\delta = \frac{(\nu - \nu_{ref})10^6}{\nu_{ref}} \tag{6.1}$$

The frequency difference is divided by ν_{ref} so that δ is a molecular property, independent of the magnetic field used to measure it. The factor of 10^6 simply scales the numerical value to a more convenient size: δ values are quoted in *parts per million*, or ppm. From an experimental point of view this referencing is extremely useful to compare the outcome of measurements coming from different instrumental setups, i.e. different applied magnetic field strength.

In previous graphs and figures, the ^{1}H and ^{13}C chemical shift have been ref-

erenced against an optimized structure of gaseous tetramethylsilane (TMS). The staggered conformation was found to be an energy minimum (T_d symmetry).

Following IUPAC recommendations [25], for aqueous solutions the primary chemical shift standard has been the methyl signal of a water–soluble derivative of TMS. In biological investigations it is proposed to adopt the methyl signal of 2,2–dimethylsilanpentane–5–sulfonic acid (DSS) at low concentration. In a dilute aqueous solution at 298 K and pH 7, the difference in ^{1}H chemicals shift between TMS and DSS amounts to 0.0173 ppm. IUPAC recommends that this difference is negligible, and data from the TMS and DSS scales may be validly compared without correction for the different ^{1}H reference [15].

In the same spirit of referencing, we propose to calculate also TMS values, in aqueous phase, according to the same sampling conditions of a given simulation protocol. This with the aim to have parameters under the same investigated computational strategy.

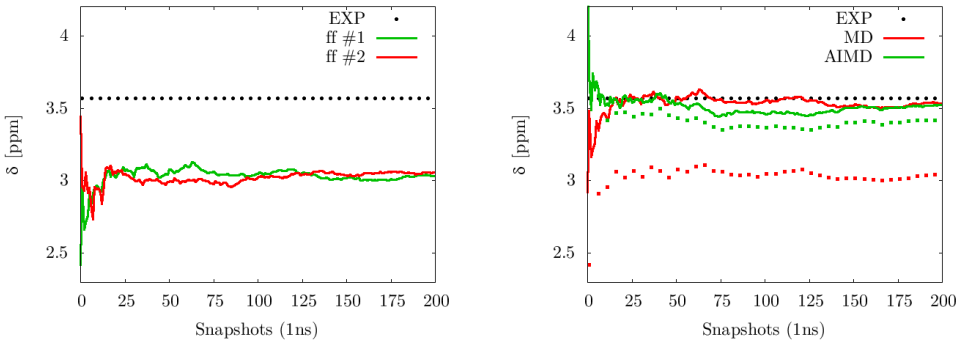
Table 6.4 summarizes the outcome of such procedure. The columns of $\Delta\delta_{\mathrm{TMS}}$ compare the values from a MD and an AIMD protocol relative to the gas phase data. As example, the deviations from TMS(gas) and aqueous MD amount up to 0.5 ppm for H and a maximum of 4.2 ppm for C.

Table 6.4: TMS reference: variation of ^{1}H and ^{13}C chemical shift depending of type of structure and sampling protocol, taking gas phase TMS as a basis of relative differences.

		$\Delta\delta_{\mathrm{TMS}}$ [ppm]	
phase		^{1}H	^{13}C
(g)	geometry optimization	0.0	0.0
(aq)	MD	0.5	4.2
(aq)	AIMD	0.1	0.9

The different TMS reference values, depending if obtained from gas phase, regular MD or AIMD, have the effect of rigidly shifting the unreferenced glycine value of different amount. When data are compared on the same

footstep, namely MD Gly(aq) with MD TMS(aq), AIMD glycine(aq) with AIMD TMS(aq), the solid line curves in Fig. (6.5b) are obtained. This procedure effectively decreases the discrepancies between calculated and experimental proton and carbon chemical shift. Gly(aq) and ethanol(aq) numerical results are reported in Table (6.5), corresponding to entries for MD[1] and AIMD[1].



(a) **(b)**

Figure 6.5: H^{α}'s floating average convergence: in Fig. (6.5a) two different force fields parametrization (ff #1 and ff #2) are used (see text for references and details), while in Fig. (6.5b) are displayed the effect of adopting different references for TMS. Dotted line (plotted every 5 steps) indicates chemical shifts relative to TMS(g) for MD or AIMD trajectories, solid lines MD– or AIMD–driven data, according to Table (6.4).

Bearing in mind the aforementioned concept to calculate the TMS parameters relying on the same approach of a given sampling protocol, QM/MM chemical shift calculations of solvated glycine are repeated also with the `g03` software package with BLYP and B3LYP functionals. These calculations however do not allow PBC to be applied, making thus possible to distinguish explicitly between the effects of finite size effect and exchange–correlation functional. The 200 configurations used in `g03` are the same of the one generated for `CP2K` in Fig. (6.5b).

The columns $\Delta\delta$(max-min) in Table (6.5) indicate the minimal and maximal difference of chemical shift of the individual atoms compared to experimental results [3, 35, 14]. Let consider classical MD–derived data only. It is to notice that for ^{13}C, `g03`/BLYP and `g03`/B3LYP isolated–system effects are a source of error up to 30 ppm and using in the calculation an hybrid

Table 6.5: Glycine(aq): effects of a CP2K/BLYP, a g03/BLYP or a g03/B3LYP MD–type of calculations, compared to an AIMD–type of approach. Ethanol(aq): MD– and AIMD–type for CP2K/BLYP.

Sampling		^{1}H		^{13}C	
		$\langle\Delta\delta\rangle$	$\Delta\delta$(max–min)	$\langle\Delta\delta\rangle$	$\Delta\delta$(max–min)
glycine(aq)	MD[1]	0.05	0.05–0.05	6.7	7.2–6.2
	MD[2]	0.08	0.12–0.03	25.5	32.6–18.4
	MD[3]	0.07	0.12–0.02	29.2	32.6–25.7
	AIMD[1]	0.04	0.04–0.04	12.6	17–8.3
ethanol(aq)	MD[1]	0.28	0.47–0.08	8.2	11.5–5
	AIMD[1]	0.11	0.18–0.04	7.0	12–1.9

[1] CP2K/BLYP, [2] g03/BLYP, [3] g03/B3LYP.

GGA functional like B3LYP, it does not alterate the situation. CP2K results are instead approximately 7 ppm apart from literature values. Data in this column are referring to C^{α} and C respectively. The $\langle\Delta\delta\rangle$ values report the average discrepancy between different atomic sites, i.e. the average of C^{α} and C. Focusing on ^{1}H column, CP2K is 0.05 ppm off and the same rationalization mentioned above for g03/BLYP and g03/B3LYP for carbon holds also for hydrogen, namely a decrease in accuracy for g03. Compared to CP2K, g03/BLYP and g03/B3LYP can differ up to 0.1 ppm per H's. The striking observation between CP2K and g03–type of calculations is that in the first case the two equivalent protons ($H^{\alpha 2}$ and $H^{\alpha 3}$) delivers the same data as expected, whilst in the second case they are 0.1 ppm apart. The atomic site equivalence is also witnessed for AIMD–derived data. Probing the equivalence of $H^{\alpha' s}$ atomic sites in CP2K supports the indication that a sufficiently large part of the intrinsic factors contributing to this symmetry are taken into account in the CP2K approach. As mentioned before, this involves the level of theoretical description, similar solvation pattern experienced of the atoms, sufficient polydiverse sampling of molecular vibrational modes, etc.

Table (6.5) resumes also the difference of CP2K MD and AIMD type of sampling results. Essentially for hydrogen the situation is unchanged, and for carbon the error is somehow increased. The conclusions for glycine(aq)

leave the overall picture unchanged in terms of performance. Alternatively, applying the same protocol for ethanol(aq), indications points towards an amelioration of the agreement with experiments in favor of the AIMD.

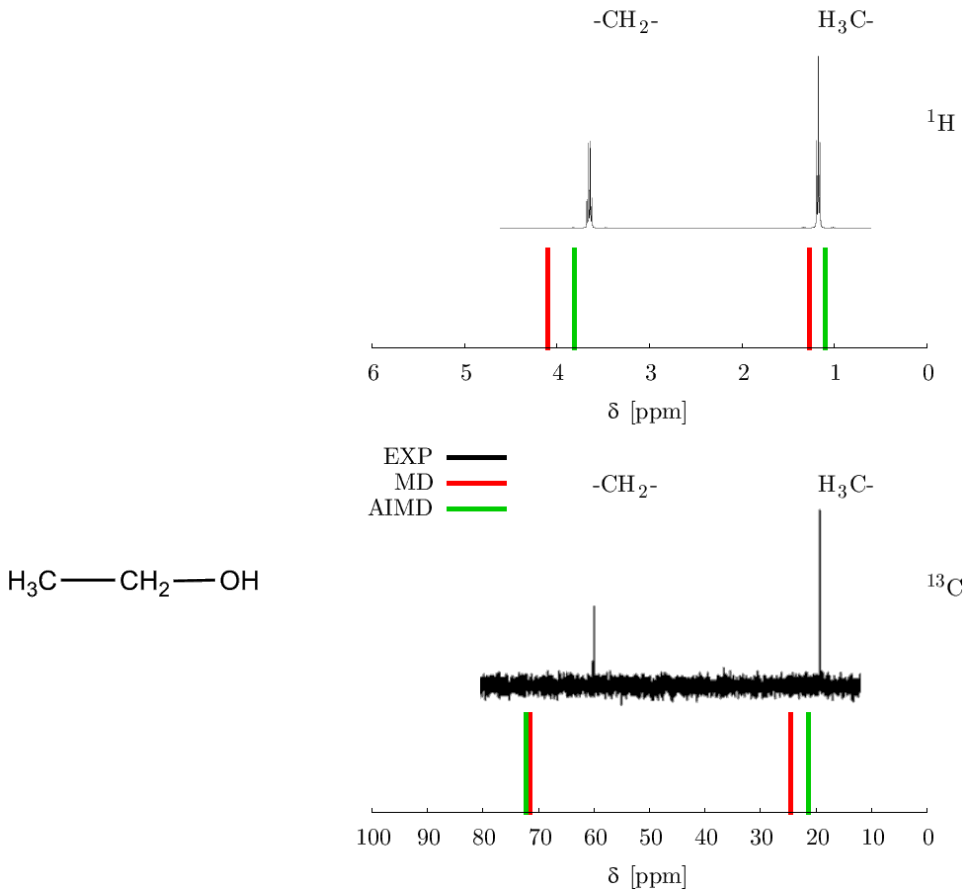


Figure 6.6: Ethanol(aq) ^{1}H and ^{13}C data: structure and representation of the experimental and calculated chemical shifts.

For a quantitative and qualitative visual appreciation of the ethanol(aq) case, in Fig. (6.6) are depicted the chemical structure as well as the experimental, MD– and AIMD–calculated chemical shift ^{1}H and ^{13}C spectra. Concerning ^{1}H experimental shift, the –CH_2– and H_3C– peaks transform themselves into a so–called quadruplet and triplet shape. This effect has to do with the lifting of the energetic level degeneracy under the effect of

spin–spin coupling, whose modeling is far beyond the scope of the present work. The actual vertical intensity of the peaks is not representative of the abundance of a given atomic specie, the integration instead represents the relative ratio of equivalent atom population. In the calculated spectra the vertical intensity is thus arbitrary. This intensity–related argumentation also holds for ^{13}C. For both atomic kinds the TMS would resonate at 0 ppm (not shown).

It is easy to see the role exerted by the hydroxyl moiety of the ethanol molecule, shifting to higher values the resonances of the –CH_2– atomic species (deshielding) compared to the less perturbed H_3C– part. The hydroxide proton in solution is not seen within the timescale of a NMR experiment because it is exchanging with other molecules, undergoing constant protonation and deprotonation processes.

6.3.2 Solvated Valine, Threonine and Tyrosine

Further studies are performed for a more complex set of amino acids. In fact glycine is the simplest representant of this category, the lateral chain being a simple proton (R = –H). Three other individual AA are investigated: valine, residue with an aliphatic chain (R = –$CH(CH_3)_2$), threonine, with the presence of an hydroxyl moiety (R = –CH(OH)–CH_3) and tyrosine, bearing a phenyl aromatic ring (R= –CH_2–Phe–OH)

The choice of these residues aims to cover different representatives of the various classes of AA. Particular the effect exerted by the presence of the hydroxyl electron withdrawing group can be assessed and validated, as well as the impact on the chemical shift induced by the aromatic ring current.

A simple MD sampling protocol, together with its parent AIMD, are applied on the three new structures, under the same settings used in the precedent examples of glycine and ethanol in aqueous phase. The only difference is the slightly increased simulation box size in order to accommodate for more water molecules due to the larger size of the solute.

In both valine and threonine cases, the calculated proton chemical shifts performed on equivalent sites indicate an high degree of convergence. Likewise to the case of glycine, the differences are in the order of a fraction of a ppm between individual atoms. These sites are the three identical protons of

individual –CH_3. The intramolecular rotation of the metyles average out the position so that every bounded hydrogen experiences a similar environment in the timescale of the trajectory simulation, both in terms of intra– as well as an intermolecular point of view.

In the case of tyrosine, instead, the situation is more complicated, since the lateral residue is a bulky phenyl ring. This large group occupies more region of space, and a ring rotation is less likely to occurs with respect to a smaller methyl group. This internal molecular motion is a low frequency mode, and the solvation pattern around this part has to change in order to accommodate the transition state during re–orientation.

An example of such structural–type of impasse is illustrated in Fig. (6.7).

Fig. (6.7a) depicts in bold blue the dihedral angle used for discussing the simulation results of tyrosine conventional MD. The chemical shifts monitored are related to the two equivalent $H^{\delta 1}$ and $H^{\delta 2}$, in the picture surrounded by the big green and red balloons respectively. Panel (6.7b) indicates that the two nuclei calculated chemical shift apparently converge in about 70 snapshots. After snapshot 100 their value starts to diverge significantly, exhibiting a sudden kink approximately at snapshot 175. Fig. (6.7d) shows individual frames δ, where the value undergoes a typical excursion during the evolution of a trajectory, while the lower panel represents the running average of these data. The running average smooths out large individual instantaneous oscillations, putting in evidence that a large deviation starts to happen around configuration 175. By looking at Fig. (6.7c), it is clear that the molecule undergoes at least two substantial structural rearrangements: the first around snapshot 75, the second around 175. By monitoring C–C^{α}–C^{β}–$H^{\delta 3}$ dihedral angle, what happens is a twisting around the C^{α}–C^{β} part of the molecule. These phenomena indicate that the atomic sites experience a different environment, hence a different chemical shift value. It is worth to mention that the magnitude of such discrepancy between nominally symmetrical atoms can rise up to 0.2 ppm, making the sampling issue perhaps the main challenge to target for increasing the accuracy of this methodological approach.

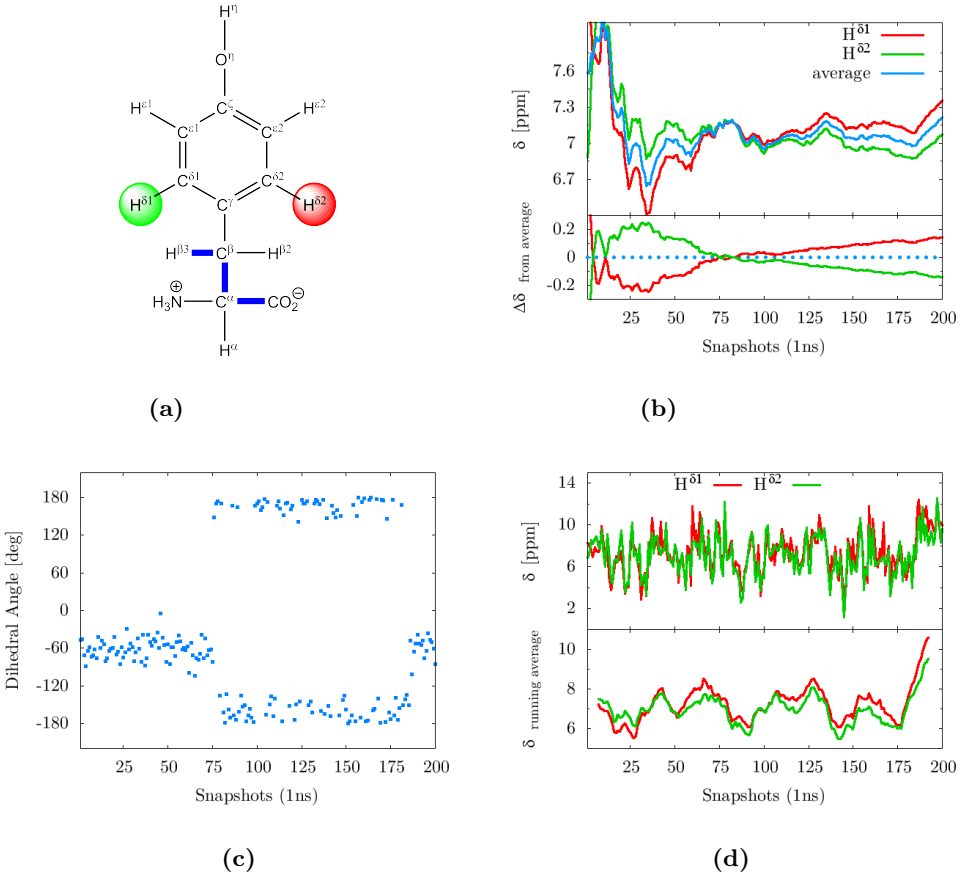


Figure 6.7: Classical MD of tyrosine(aq): example of a multi–state structure. In Fig. (6.7c) is monitored the value of the C–C^{α}–C^{β}–$H^{\delta 3}$ dihedral angle (bold blue lines in Fig. (6.7a)). Fig. (6.7b) displays incremental average chemical shift for $H^{\delta 2}$ and $H^{\delta 3}$ atoms while sampling a standard MD trajectory. The upper panel in Fig. (6.7d) displays the chemical shifts of the two protons along the sampled snapshots, the lower one illustrating their running average of 13 points window.

The variations of ^{1}H δ value coming from basis set expansion, level of theory and extension of the quantum region considerations, in fact accounts for several 10^{-2} ppm, whilst the discrepancy between MD and AIMD and the sampling problem do exerts an impact in the 10^{-1} magnitude.

In the forthcoming part the bottleneck of accurate sampling will be targeted.

Enhanced Sampling – Hamiltonian Replica Exchange

Conformational sampling is an essential concern to the study of complex molecular systems. A major obstacle for the correct sampling of such systems is the fact that the potential energy surface can be very rugged and contains a large number of local energy minima.

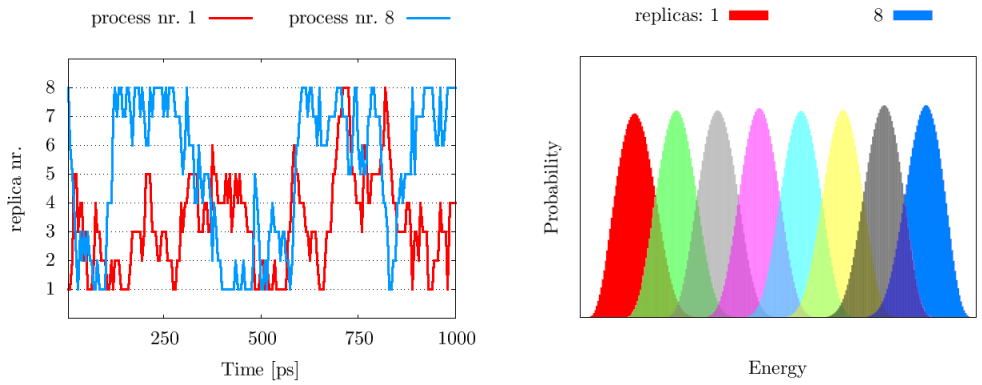
A conformation has to be taken as initial guess and serves as a basis of a MD simulation. The paradox is that even if this starting configuration is energetically higher than another structure, depending on the shape of the potential energy surface, the latter may rarely or never be observed. This depends on the energetic barriers separating the states and the length of the simulation. This cause kinetic trapping due to low crossing rates.

We speak in this case to "accelerate" the MD and make the conformational sampling more efficient than a conventional MD strategy.

Indeed the potential problem with standard approaches is that energetic barriers higher than the thermal kinetic energy of the system are rarely overcome. This means that certain states or configurations that are normally populated at a given temperature, but separated by barriers in the order of several k_BT, are not easily accessible within the timeframe of the simulations. To cure this problem one needs to accelerate this "rare events" occurrence, namely to cross those energetic separations in order to perform more realistic sampling. In this cases one have to rely on the so–called class of "enhanced sampling techniques" and it is decided here to adopt the strategy of replica exchange.

The replica exchange method (REM) is based on multiple concurrent (parallel) canonical simulations that are allowed to occasionally exchange their configurations. Configurations between couple of replicas are tentatively exchanged at prescribed time intervals using a Metropolis–like probabilistic criterion. Typically one sets increasing temperature for each replica: the target temperature, i.e. the temperature corresponding to the thermodynamic state of interest, is usually the lowest among all replicas. In this manner, configurations from "hot" replicas, i.e. configurations where energy barrier are easily crossed, may be occasionally accepted as the target temperature. The global state of the extended system may evolve in two ways: by evolving each replica independently (via MD simulation protocols) and

by exchanging the configurations of two replicas. The sampling in the multi–configurational space in REM evolves towards a global equilibrium defined by the multi–canonical probability distribution of the extended system. The condition is that each parallel process must perform a random walk in the temperature domain.



(a) **(b)**

Figure 6.8: (6.8a) Hamiltonian REM Simulation with 8 replicas of tyrosine (aq): path followed by two processes along swapping MD runs. To reconstruct a trajectory for a given replica, data from several processes have to be combined. (6.8b) Configurational energy distribution for 8 replicas: The overlap between two distribution is a lower bound for the acceptance probability.

In many cases, one has to effectively sample coordinates that are rather localized in the system (e.g. the solute). In normal temperature REM, the heat in the hot replicas is clearly distributed among all the degrees of freedom of the system and therefore most of this heat is wasted for exchanging uninteresting configurations (e.g. bulk solvent configurations). The algorithm we adopt is a variant of the replica exchange called Hamiltonian REM, that is far more flexible than standard temperature REM technique described above. In the Hamiltonian REM, each replica is characterized by a different potential energy rather than by a temperature. In its simplest implementation, the classical potential energies of the replicas differ by a scaling factor c_m, with $c_1 = 1$ for the target replica (replica nr. 1 in Fig. (6.8). Another advantage of using the Hamiltonian REM is that as all the replicas have the same operating temperature, one does not have, like in temperature REM, to reinitialize

or rescale the velocities after one successful configuration exchange.

In this computational framework the sampling of phase space is enhanced by the mean of a regular MD run and a series of artificial modification of the potential energy on the solute only. Several parallel MD simulations with different potential energy surfaces are simultaneously carried out, and configurations are allowed to swap at fixed intervals of time. The goal is to collect the data generated for the unperturbed trajectory which incorporate configurations coming from perturbed one, where energetic barriers are easily crossed compared to the unmodified potential one only.

Hamiltonian REM classical MD calculations are performed using the `ORAC` software package [26]. Swaps are attempted every 1000 steps, and the acceptance ratio is around 15–20 % of the attempts.

Furthermore, we can also for REM–MD propagate on a given snapshot a corresponding *ab initio* MD, accessing finally to REM–AIMD type of calculations.

For consistency also glycine(aq) calculations have been performed according to this simulation protocol, however there is no gain in accuracy: the new results matches almost identically the previous results (data not shown). Concerning aqueous valine, threonine and tyrosine, the results are summarized in Table (6.6).

Simulated relative ^{1}H chemical shifts according to pure MD sampling indicate an unsigned discrepancy compared to the experimental data [35] of all individual protons ranging from 0.8 to 0.0 ppm. The upper bound of 0.8 ppm is indicated by the highest value in the $\Delta\delta$(min–max) column for every MD method. The average error $\langle\Delta\delta\rangle$ is in two case half of this maximal value. This parameter indicates the spread of the inaccuracy of the method. A slight gain in accuracy is obtained relying on AIMD. The maximal error decreases by 0.2 ppm in two cases and the spread of the errors consequently. Concerning ^{13}C MD results performance ranges between 14 and roughly 3 ppm. For this atomic kind, AIMD trajectory ameliorate the refinement in some cases and simultaneously increases the error in others, leaving unchanged the average error but for threonine.

By comparing MD and REM–MD results, we see a systematic downsize of the discrepancy between simulated and literature data, this for both atomic

Table 6.6: valine(aq), threonine(aq) and tyrosine(aq): effects of a standard and a REM–type of sampling, both for MD and AIMD.

		^{1}H		^{13}C	
Sampling		$\langle\Delta\delta\rangle$	$\Delta\delta$(max–min)	$\langle\Delta\delta\rangle$	$\Delta\delta$(max–min)
valine	MD	0.6	0.7–0.4	7.2	14–3
	AIMD	0.3	0.5–0.1	7.8	20–0.9
	REM–MD	0.1	0.3–0.0	6.2	9.3–3.2
	REM–AIMD	0.3	0.3–0.2	5.4	14.3–0.5
threonine	MD	0.3	0.8–0.0	8.3	12.9–3.5
	AIMD	0.2	0.3–0.0	11.3	20.7–3
	REM–MD	0.2	0.4–0.1	6.0	10.4–3.6
	REM–AIMD	0.2	0.5–0.0	7.8	15.9–0.4
tyrosine	MD	0.4	0.8–0.0	7.0	14.5–1.5
	AIMD	0.3	0.8–0.0	6.3	16.7–0.4
	REM–MD	0.1	0.3–0.0	7.0	14.2–3.2
	REM–AIMD	0.2	0.5–0.1	5.6	12.6–0.5

species. Replica exchange decreases the maximum error and the value of the average by a factor of two for hydrogen and 1/3 for carbon. The same is also true between AIMD and REM–AIMD data, although the amelioration is to a smaller extent.

Hydrogen data for REM–MD and REM–AIMD sampling leaves the picture unchanged, indicating that the two methods perform equally well, the average accuracy being between 0.1 and 0.3 ppm. This comment does not apply to carbon: the maximum error increases in two cases out of three, but on the other hand displays opposite tendency: the minimal error approaches 0.5 ppm. The REM–AIMD error average is e few ppm below the half of the maximum discrepancy, indicating that these maxima could be affected by a systematic error. Except for the last case, the general trend is that the $\langle\Delta\delta\rangle$ represent approximately 1/2 of $\Delta\delta$(min–max), indicating even balancing between under and overestimated data deviation from experiment.

An assessment of the accuracy of the sampling method is performed via

calculation of the standard deviation coming from averaging process. Considering the 95 % of the confidence interval, the typical statistical error of this type of sampling is estimated to be ± 0.2 and ± 1 ppm for H and C respectively.

Fig. (6.9) details the molecular structure of tyrosine, as well as presenting the comparison of literature and the REM–AIMD calculated NMR chemical shift spectra of ^{1}H and ^{13}C of this AA in water.

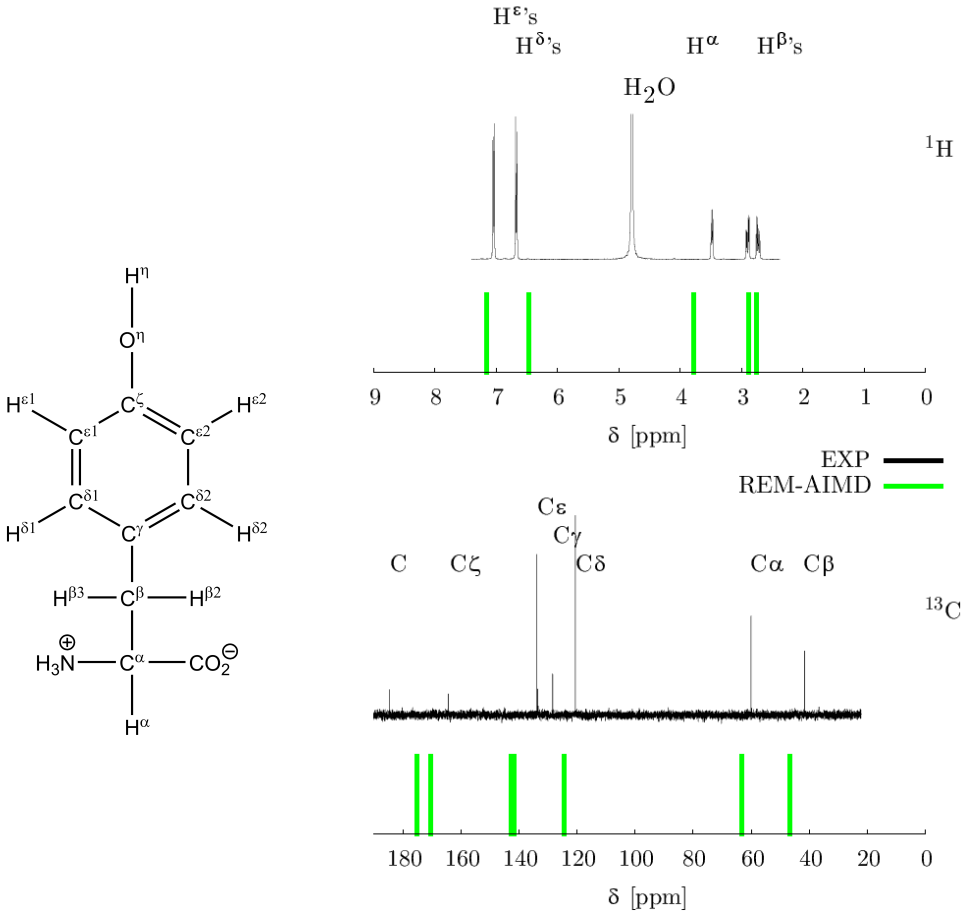
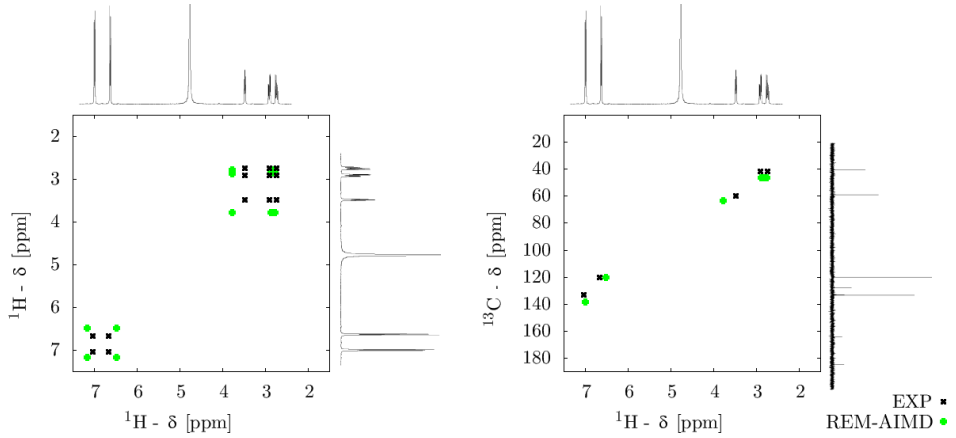


Figure 6.9: Tyrosine ^{1}H and ^{13}C data: structure and representation of the experimental and REM–AIMD simulated chemical shifts.

Two–Dimensional NMR

From an experimental point of view, over the last three decades, two–dimensional (2D) methods have revolutionized the practice of NMR spectroscopy. Selective 2D NMR pulse sequences induces complex magnetization components on a site–specific level, producing 2D spectrum as a function of two chemical shift variables. Examples of such 2D experiments are total correlation spectroscopy (TOCSY) and heteronuclear single quantum coherence (HSQC) spectroscopy. In both techniques magnetization is transferred selectively between adjacent nuclei. The efficiency of this intensity transfer between spins is related to the proximity of such sites. For instance, TOCSY and HSQC spectra thus serves to identify all pairs of nuclei which are in the vicinity or chemically bounded respectively.

Combining the results of the previous section, 2D spectra can be obtained. The outcome of such procedure for Tyrosine is depicted in Fig. (6.10). The same deviation from experimental data of the 1D spectra affects the 2D results.



(a) 2D [H,H]–TOCSY **(b)** 2D [H,C]–HSQC

Figure 6.10: Tyrosine correlation spectroscopy: total correlation spectroscopy (TOCSY) and heteronuclear single quantum coherence (HSQC) spectroscopy data from REM–MD.

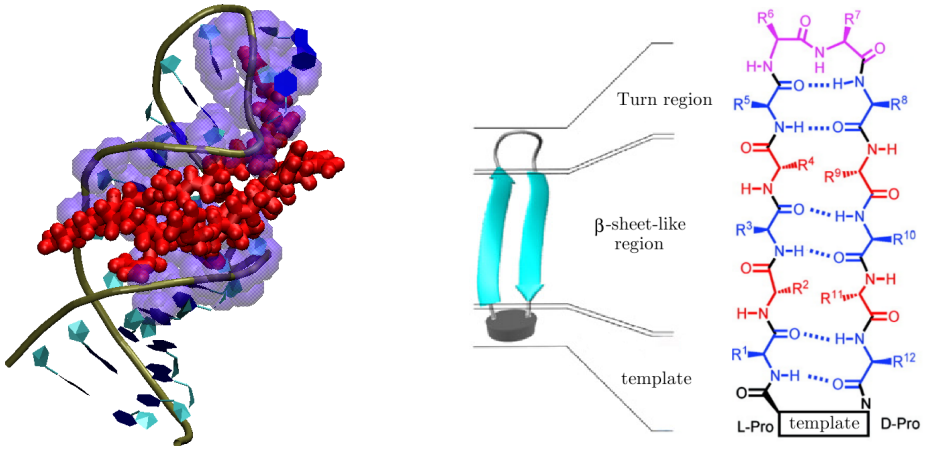
6.4 Solvated Polypeptide: β–Hairpin BIV2

Increased structural complexity arises by considering a molecule formed by two or more amino acids linked together via the peptide bond. An important aspect to underline, due to the floppy conformation that can be adopted by polypeptides, is that when there is a favorable intermolecular interaction energy between different moieties, supramolecular structures are created by folding the macromolecule into regular structures. Notable examples of such kind of arrangements are the α–helix, the β–sheet and the disulfide bond.

As a case study, we aim to simulate NMR chemical shift parameters for a peptide known to adopt a stable conformation. BIV2, a so–called peptidomimetic ligand for bovine immunodeficiency virus (BIV) Tat protein, that inhibits binding to the transactivator response element (TAR) RNA. BIV2 is derived by grafting onto a hairpin–inducing D–Pro–L–Pro template a sequence related to the natural viral RNA recognition element in Tat [8]. This conformationally constrained peptide adopt a β–hairpin structure. The transcriptional activator protein (Tat) plays a critical role in the viral life cycle and contribute to the proliferation of the virus.

The 3D structure of BIV2–TAR complex has been solved by multidimensional heteronuclear and nuclear Overhauser enhancement spectroscopy NMR methods. This model is represented in Fig. (6.11a), with BIV2 as red ball and sticks, binding into a groove of RNA.

Fig. (6.12) shows the side view of the BIV2 14–residues peptide. The residue sequence *in extenso* is cyclo–(D–Pro–L–Pro–Arg^{1}–Val^{2}–Arg^{3}–Thr^{4}–Arg^{5}–Gly^{6}–Lys^{7}–Arg^{8}–Arg^{9}–Ile^{10}–Arg^{11}–Val^{12}). The D–Pro–L–Pro template linkage introduces an element of rigidity in the cyclic peptide, framing thus the rest of the loop in such a way to promote intramolecular H–bonds of the peptide mainchain. The protein backbone skeleton lies approximately in a plane. In the β–type of secondary structure, lateral chains in the sequence are alternatively pointing up and down. In the side view shown in Fig. (6.12) , the sidechain in blue are pointing all in the upwards direction, while the red downwards. The first include four Arg residues, making this part of the molecule highly polar. In the BIV2–TAR RNA structure, residue in this region (blue) are actively participating in the protein–RNA interacting within the epitope region (the target of an immune response), whilst the



(a) BIV2–TAR complex **(b)** BIV–frame

Figure 6.11: BIV2 cyclic peptide upon binding to the TAR RNA (PDB entry 2A9X). A schematic representation of the 12–residues loop mounted on the D–Pro–L–Pro template is shown on the right.

Figure 6.12: β–hairpin structure of BIV2.

opposite face of the molecule (red) is solvent exposed.

Coordinates for the BIV2 peptide–TAR RNA complex are deposited in the Protein Data Bank (PDB) [2] as entry 2A9X. From an experimental point of view, NMR spectra were recorded typically at a peptide concentration of 10–20 mg/mL.

The starting configuration of the simulation is the BIV2 peptide where the RNA has been deleted from the coordinates. We aim to run conventional MD and REM–MD, calculating chemical shifts on these trajectories only, always adopting a QM/MM type of scheme, where the QM region incorporate the solute and the MM the solvent. Same computational parameters are used as in the amino acids studies carried out before.

According to the two different sampling protocols, two different trajectory lengths are obtained: a dynamics of 2 ns for MD and a 3 ns for REM. Relative chemical shifts averages are calculated on the whole production trajectory for MD, while only the last 2.5 ns are used for REM. The first 0.5 ns sought for equilibration when switching on the REM method. In this first part of the trajectory the acceptance criterion of the swapping probability, used as an estimate for algorithmic convergence, is not yet oscillating around an equilibrium value (data not shown). This criterion is an indicator of the stability of the method. The production trajectory gives rise to 400 individual structure calculations for each nucleus during MD, and 500 for REM.

For structural interpretation of the whole subset of trajectories, the definition of the "Dictionary of protein secondary structure" (known as DSSP) proposed by Kabsch and Sander [18] is used. This is based on a pattern–recognition process of hydrogen–bonded and geometrical features, that allows to recognize elements of 3D structure.

In a nutshell, the DSSP recognition algorithms exploits the presence or absence of an H–bond between mainchain N–H$\cdots$O triplets; "n–turns" with an H–bond between the C=O of residue i and the NH of residue $i+n$, where $n = 3, 4, 5$, and "bridges" with H–bonds between residues not close to each other in sequence. Repeating 4–turns define α–helices, and repeating bridges define β–structure.

Fig. (6.13) illustrates graphically the elements of DSSP observed for every single of the 14 residues of the peptide along a given trajectory. The lower panel represents the results based on the snapshots of MD, in the upper part the 8 replicas of the REM. The ordinate axis of each strip of data indicate the residue number. Residue 13 and 14 are the D–Pro and L–Pro respectively.

Focusing on the 2 ns MD trajectory, it can be noticed that residue 2 and 11 participate in a B–bridge type of structure. A second characterizing structural element corresponds to a bend, shared between residues 5, 6, 7 and 8. Bends are regions with high curvature. The invoked H–bonds are kept in place during the whole trajectory. Substantially the MD trajectory shows a static picture, where the same steady–state structure is monotonously preserved throughout the whole range.

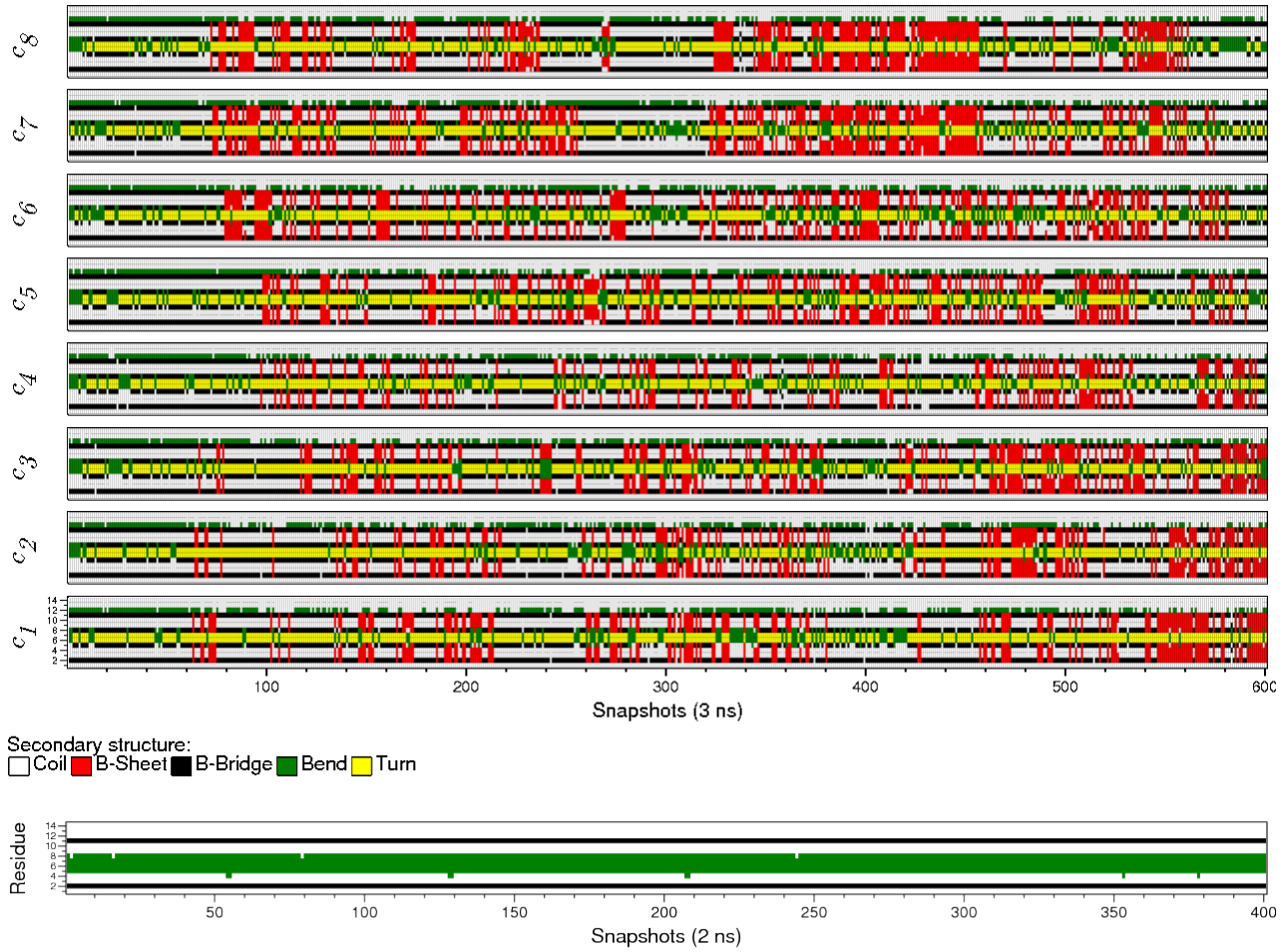


Figure 6.13: Dictionary of protein secondary structure: in the lower part the MD trajectory, in the top panels the 8 replicas, indicated by the corresponding c's.

Very different is the case of the REM trajectory: considering the target replica indicated by c_1, since the very beginning of the simulation it is assisted at the modification and the introduction of novel structural arrangements. In particular, to the addition of the aforementioned features of classical MD, the establishment of an extended turn (yellow), which persist as long as the simulation progresses. Eventually even a bridge conformation is observed (red), extending over residues 2–5 and 8–11 of the molecule. This latter is transiently created. As a consequence of the REM swapping, a scrambling of structures is observed, replica c_1 targeting the more stable ones.

(a) MD **(b)** REM

Figure 6.14: Backbone representation of BIV2 structures. Solid lines representing stable interactions, dots are the observed transiently. Color code taken from Fig. (6.13)

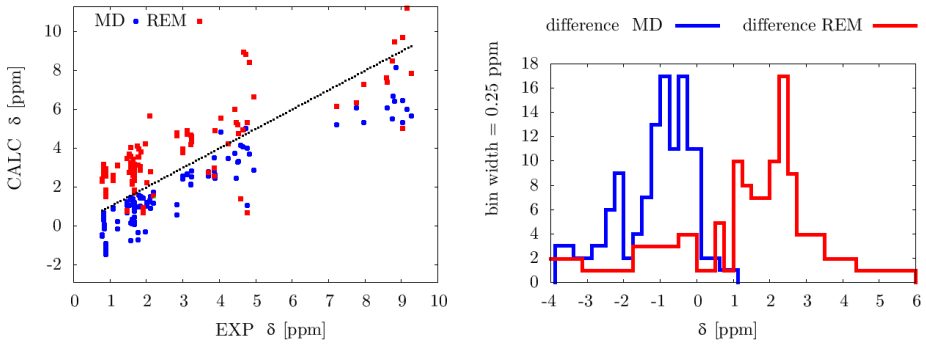
Fig. (6.14) summarizes the nature and the persistence of the different types of interactions encountered during MD and REM simulations. The same color code is in Fig. (6.13) is adopted. The reason of the apparent staticity of the MD trajectory compared to the diversity of morphologies experienced while performing REM is rooted in the different capability of cross activation of barriers discussed before, allowing the further stabilization of other intramolecular secondary elements.

As already said, the starting point for both simulations is taken from a BIV2–TAR deposited structure, where the peptide is complexating the RNA. Experimental evidence indicates that the binding mode of the substrate does not involve any particular structural rearrangement compared to the form in solution [23] This could be true for the already constrained backbone conformation, but there is no information concerning the change in orientation of the lateral chains.

The unbound peptide, instantaneously freed from its counterpart, could probably have not had sufficient time to cross the energetic barriers preventing it to adopt a more stable conformation in solution. The simple fact of

deleting RNA coordinates, and using it as initial guess for the simulation, the BIV2–alone could have stuck this conformation in a local minimum difficile to escape from. This is not the case for the REM trajectory, where energetic barriers are artificially lowered, allowing the molecule to sample extended region of phase space.

The outcome of the chemical shifts calculations is assessed in Fig. (6.15). The data represents calculations performed on 101 hydrogen atoms present in the peptide averaged over 400 and 500 snapshots, corresponding to the MD and REM trajectories respectively.



(a) individual differences

(b) histogram of differences

Figure 6.15: Left panel: experimental *vs* calculated hydrogen chemical shift. Dotted line represents the identity as expressed by CALC = EXP. Left panel: histogram of differences between calculated and experimental values.

Fig. (6.15a) indicates the datapoints and the straight line $f(x) = x$: the latter being the ideal correlation between observed and theoretical values. The results obtained from MD sampling are systematically lower in estimating the chemical shifts (blue points), whilst the REM–MD are more scattered symmetrically around the experimental values (red points). The distribution of the errors in Fig. (6.15b) indicates a narrower spread in favor of the MD–obtained data by respect of the REM–MD, although from a sort of bimodal profile.

The discrepancy of the calculated values compared to the measured ones, lies in the order of few units of ppm, and its dispersion extends over a large

scale, resulting to a qualitative agreement between the results of this work compared to the experiments. The presented methodology indicates that calculation of chemical shifts of full proteins in solution is indeed possible, even if the results are not yet on a predictive level. This because many sources of errors cannot be controlled or isolated. Besides possible inaccuracies of the quantum chemical model, i.e. coming from the used density functionals and inherent to the basis set convergence, the timeframe of the simulated trajectory could affect the reliability of the ensemble average. In this respect the examined structures are poorly representing experimental equilibrium conditions. This can be especially true for NMR spectroscopy, where the experimental timescales are in the order of the millisecond. Slow modes sampling would need to extend the molecular dynamics beyond the limits of the nanosecond investigated in this work. Another contribution to the discrepancy could be related to the accuracy of the geometries. In analogy to the case of isolated amino acids, AIMD is expected to provide a remedy to this aspect. Furthermore it would permit to sample fast modes, where intrinsic QM effects could play an important role. An alternative way out of this dilemma is to divide the whole collection of structures into selected subsets of structures sharing common features.

Clustering MD Configurations

The paradigm of the previous sections can be summarized in the concept that the whole set of structures are contributing equally to the ensemble average. Results based on this assumption would assign an equal weight to every snapshot, and the mean value of the whole dataset can be used for comparison to the experimental data. In some cases, one wants to identify states which are more frequently and repeatedly populated, regardless of when they occur in a simulation. This is the field of statistical cluster analysis. It is decided here to adopt the algorithm proposed by Torda and van Gunsteren [34, 7]. The procedure is briefly described hereafter. To find clusters of structures in a trajectory, the root mean square difference (RMSD) of atomic cartesian coordinates between all pairs of structures is determined. For each structure the number of other structures for which the RMSD is lower than a given threshold is calculated (set to 0.05 nm). The conformation with the highest number of neighbors is taken as center

of a cluster, and formed together with all its neighbors a cluster (Cluster 1). The conformations of this cluster are thereafter eliminated from the pool of structures and the process is repeated until the pool of structures is empty. In this way, a series of non–overlapping clusters of structures is obtained. Considering the 500 REM geometries, the sizes of the most three populated clusters have 78, 41, and 33 members respectively.

A schematic view of the procedure of such clustering procedure is summarized in Fig. (6.16), indicating structures belonging to a given cluster. This is an individual RMSD plot corresponding to snapshot nr. 100. Again the data of the first 0.5 ns are discarded for equilibration purposes. RMSD are calculated after taking the precaution of performing a backbone alignment fitting; minimizing thus the effects of rotational tumbling of the molecule in solution.

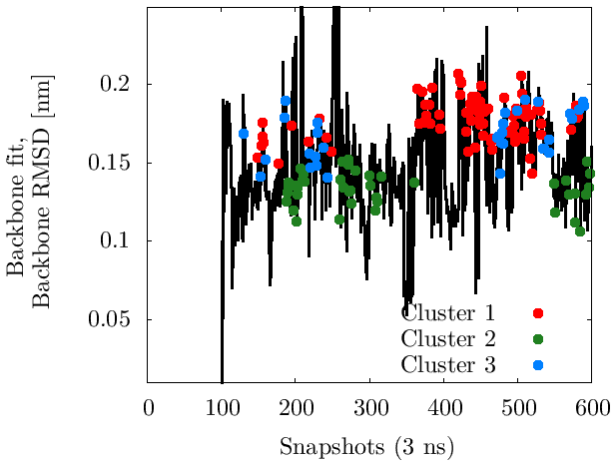


Figure 6.16: Backbone RMSD of atom positions of the simulated structures from the REM trajectory.

2D–NMR permits to highlight the selective interactions between pairs of selected atoms, or groups of chemical–shift–equivalent atoms. 2D [H,N]–HSQC, representative of the N–H moiety, are calculated for all residues and are presented in Fig. (6.17).

The spectral information is spread into two frequency dimensions and allows simultaneous collection of information about resonances, thus improving the effective resolution and sensitivity [38]. Useful information are contained in peaks which correlate the frequencies of spins that are allowed to exchange

information via magnetization transfer.

The chemical shifts of ^{15}N are obtained by indirect referencing, i.e. referenced to TMS standard, using conversion factors derived from ratios of NMR frequencies [25, 39]. For example, the zero frequency for ^{15}N is obtained from the calculated TMS ^{1}H by multiplying the latter by the so–called ^{15}N/^{1}H nucleus–specific Ξ ratio, corresponding to 0.1013.

In the example of regular MD, the hydrogen values of the 2D [H,N]–HSQC are contained in a narrow hydrogen band interval. This 2D spectrum is indicative of the strength of a given intramolecular N–H$\cdots$O hydrogen bond. When lacking of secondary structural elements, the peptide tends to a conformation known as the random coil limit. The crucial point is that in such unstructured, flexible forms, identical amino acids constituting the peptide experience similar environment. This is due to the absence of specific interactions shaping a characteristic immediate surroundings. In other words, in the random coil limit, same amino acids within a protein display the same chemical shift. In fact, for N–H protons, experimentally these value ranges from 8.1 (Trp) to 8.75 ppm (Asn) [40].

The 2D [H,N]–HSQC average of the clusters is depicted in Fig. (6.17). The lower part presents the results for the classical MD, where the chemical shift for hydrogen values spans over 3.5 ppm. Two valine residues are nr. 2 and 12, and their position in the 2D spectrum are very close (see approximately coordinates [8:122] ppm). This observation can be rationalized by the assumption that all amino acid side chains in an extended, flexible polypeptide chain are exposed to the same solvent environment, so that multiple copies of a specified amino acid in the sequence have identical chemical shifts. Although the chemical shift is primarily determined by the covalent structure of the amino acid residue, it can be also be significantly affected by the interaction with the solvent. Therefore, the exclusion of the solvent water from the interior of the loop causes the chemical shifts of the N–H to be different from those of the water–exposed residues. This conformation–dependent chemical shift dispersion is found to be sufficiently large to enable ^{1}H NMR studies of protein denaturation [40].

For the REM clusters the inter–chain N–H$\cdots$O hydrogen–bond organizes the 3D complex structure (more compact), so that spatial folding of the peptide chain is manifested in the chemical shift, resulting in a dispersion

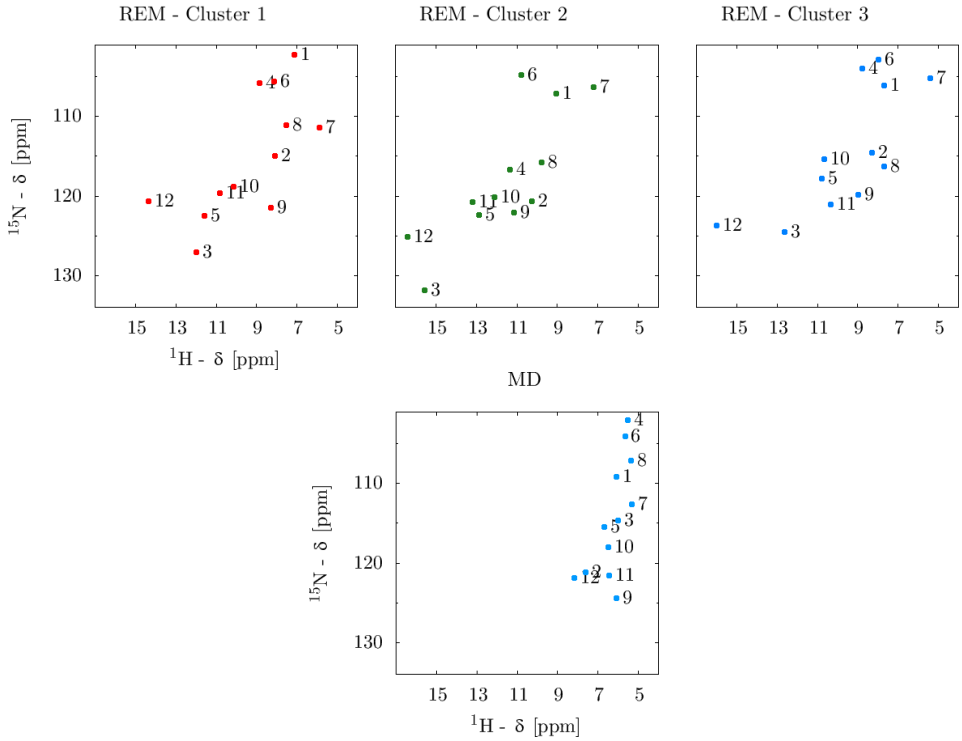


Figure 6.17: 2D [H,N]–HSQC.

(broadening) of the shifts relative to random coil reference data (see upper three diagrams of Fig. (6.17)).

6.5 Conclusions

With its roots in physics, the NMR frequencies and correlations of the signals reveal fundamental chemical and structural properties of molecules. In the context of computer simulations, the ultimate goal is to characterize and predict the behavior of real systems. Whether this target is achieved depends both on the quality of the model used and on a trade off with the available computational resources. Recognized that no model is an exact representation of the real system, the adopted model must be verified by comparison to experimental data.

In this work it is decided to sample configurational space via molecular dynamic approaches. Macroscopic properties are ensemble averages on a rep-

resentative statistical ensemble of molecular systems. By averaging over an equilibrium trajectory, properties can be extracted. It is therefore necessary to include conformational averaging, where direct effects due to changes in the populations of the conformers contribute to the averaged NMR spectrum or due to a solvent effect through interactions with the solute molecules.

Ab initio derived trajectories appear to better capture the oscillations of the fastest degrees of freedom, in particular involving C–H bonds. Validity and accuracy of the proposed approaches have been assessed by using structure verification procedures, in conjunction to 1D NMR spectra and 2D NMR correlation spectra.

The presented simulation methods aim to study processes in which a wide range of chemical environments are sampled. This investigation indicate the prospect of using biomolecular NMR for detailed studies like protein folding, in particular for distinguish between two–state and multi–state folding and unfolding transitions. Finally to obtain information about physiologically active, folded forms of proteins.

References

[1] A.D. Becke. Density–functional exchange–energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098, 1988.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235, 2000.

[3] Biological Magnetic Resonance Data Bank. http://www.bmrb.wisc.edu, 2012.

[4] Rosa E. Bulo, Christoph R. Jacob, and Lucas Visscher. Nmr solvent shifts of acetonitrile from frozen density embedding calculations. *J. Chem. Phys. A*, 112:2640, 2008.

[5] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, et al. AMBER 9. *University of California, San Francisco*, 2006.

[6] A. Cavalli, X. Salvatella, C.M. Dobson, and M. Vendruscolo. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U.S.A.*, 104:9615, 2007.

[7] X. Daura, K. Gademann, B. Jaun, D. Seebach, W.F. van Gunsteren, and A.E. Mark. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.*, 38:236, 1999.

[8] Amy Davidson, Krystyna Patora-Komisarska, John A. Robinson, and Gabriele Varani. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.*, 39:248, 2011.

[9] J.N. Dumez and C.J. Pickard. Calculation of NMR chemical shifts in organic solids: accounting for motional effects. *J. Chem. Phys.*, 130:104701, 2009.

[10] H. Flyvbjerg and H.G. Petersen. Error estimates on averages of correlated data. *J. Chem. Phys.*, 91:461, 1989.

[11] R. Freeman. *Spin Choreography.* Oxford University Press, New York, 1998.

[12] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[13] S. Gnanasambandam, Z. Hu, J. Jiang, and R. Rajagopalan. Force field for molecular dynamics studies of glycine/water mixtures in crystal/solution environments. *J. Phys. Chem. B*, 113:752, 2008.

[14] H.E. Gottlieb, V. Kotlyar, A. Nudelman, et al. NMR chemical shifts of common laboratory solvents as trace impurities. *J. Org. Chem.*, 62:7512, 1997.

[15] R.K. Harris, E.D. Becker, S.M. Cabral De Menezes, R. Goodfellow, and P. Granger. NMR nomenclature: nuclear spin properties and conventions for chemical shifts. IUPAC recommendations 2001. International Union of Pure and Applied Chemistry. Physical chemistry division. Commission on molecular structure and spectroscopy. *Mag. Res. Chem.*, 40:489, 2002.

[16] J. Hutter and D. Marx. Proceeding of the february conference in Jülich. In J. Grotendorst, editor, *Modern methods and algorithms of quantum chemistry*, Jülich, 2000. John von Neumann Institute for Computing.

[17] P.K. Janert. *Data Analysis with Open Source Tools.* O'Reilly Media, 2011.

[18] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen–bonded and geometrical features. *Biopolymers*, 22:2577, 1983.

[19] R. Kaptein, R. Boelens, RM Scheek, and WF Van Gunsteren. Protein structures from NMR. *Biochemistry*, 27:5389, 1988.

[20] M. Kaupp, M. Bühl, and V.G. Malkin. *Calculation of NMR and EPR parameters: theory and applications*. Wiley–VCH, 2004.

[21] F.G. Klärner, B. Kahlert, A. Nellesen, J. Zienau, C. Ochsenfeld, and T. Schrader. Molecular tweezer and clip in aqueous solution: unexpected self–assembly, powerful host–guest complex formation, quantum chemical 1H NMR shift calculation. *J. Am. Chem. Soc.*, 128:4831, 2006.

[22] C. Lee, W. Yang, and R.G. Parr. Development of the Colle–Salvetti correlation–energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785, 1988.

[23] Thomas C. Leeper, Zafiria Athanassiou, Ricardo L. A. Dias, John A. Robinson, and Gabriele Varani. TAR RNA recognition by a cyclic peptidomimetic of Tat protein. *Biochem.*, 44:12362, 2005.

[24] G. Magyarfalvi and P. Pulay. Assessment of density functional methods for nuclear magnetic resonance shielding calculations. *J. Chem. Phys.*, 119:1350, 2003.

[25] J.L. Markley, A. Bax, Y. Arata, CW Hilbers, R. Kaptein, B.D. Sykes, P.E. Wright, and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.*, 280:933, 1998.

[26] S. Marsili, G.F. Signorini, R. Chelli, M. Marchi, and P. Procacci. ORAC: a molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. *J. Comput. Chem.*, 31:1106, 2010.

[27] S. Moon and D.A. Case. A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: mixed basis set and ONIOM methods combined with a complete basis set extrapolation. *J. Comput. Chem.*

[28] C.J. Pickard and F. Mauri. All–electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B*, 63:245101, 2001.

[29] D. Sebastiani and M. Parrinello. A new ab–initio approach for NMR chemical shifts in periodic systems. *J. Chem. Phys. A*, 105:1951, 2001.

[30] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J.M. Aramini, G. Liu, A. Eletsky, Y. Wu, K.K. Singarapu, A. Lemak, et al. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.*, 105:4685, 2008.

[31] S.G. Spanton and D. Whittern. The development of an NMR chemical shift prediction application with the accuracy necessary to grade proton NMR spectra for identity. *Mag. Res. Chem.*, 47:1055, 2009.

[32] JF Stanton, J. Gauss, ME Harding, and PG Szalay. `CFOUR`, a quantum chemical program package. *For the current version, see http://www.cfour. de.*

[33] L. Szilágyi. Chemical shifts in proteins come of age. *Prog. Nucl. Magn. Reson. Spectrosc.*, 27:325, 1995.

[34] A.E. Torda and W.F. van Gunsteren. Algorithms for clustering molecular dynamics configurations. *J. Comput. Chem.*, 15:1331, 2004.

[35] E. L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R. Kent Wenger, H. Yao, and J.L. Markley. Biological Magnetic Resonance Data Bank. 36:D402, 2008.

[36] J.A. Vila, J.M. Aramini, P. Rossi, A. Kuzin, M. Su, J. Seetharaman, R. Xiao, L. Tong, G.T. Montelione, and H.A. Scheraga. Quantum chemical 13Cα chemical shift calculations for protein NMR structure determination, refinement, and validation. *Proc. Natl. Acad. Sci. U.S.A.*, 105:14389, 2008.

[37] Valery Weber, Marcella Iannuzzi, Samuele Giani, Jurg Hutter, Reinout Declerck, and Michel Waroquier. Magnetic linear response properties calculations with the gaussian and augmented–plane–wave method. *J. Chem. Phys.*, 131:14106, 2009.

[38] D.E. Wemmer and B.R. Reid. High resolution NMR studies of nucleic acids and proteins. *Annu. Rev. Phys. Chem.*, 36:105, 1985.

[39] D.S. Wishart, C.G. Bigam, J. Yao, F. Abildgaard, H.J. Dyson, E. Oldfield, J.L. Markley, and B.D. Sykes. ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, 6:135, 1995.

[40] K. Wüthrich. *NMR of Proteins and Nucleic Acids.* Wiley, New York, 1986.

[41] K. Wüthrich. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angew. Chem. Int. Ed.*, 42:3340, 2003.

Acknowledgments

At first, I would like to express my gratitude to Prof. Dr. Jürg Hutter for the support, in its broadest sense, displayed in supervising this Thesis. Thanks to him I could practice a PhD experience in absolute academic freedom.

I am particularly thankful to Dr. Teodoro Laino, Dr. Marcella Iannuzzi–Mauri and Dr. Valéry Weber.

Working with these magnificent four, I could benefit from their guidance, encouragement, and concrete contributions regarding both the scientific aspects of this work, as well as enjoying the daily life with positive attitude, rich of humor too.

The `CP2K` community has to be mentioned for the good spirit of collaborative support that characterizes the open source philosophy.

Members of the Promotionskomitee's work shall also be mentioned here.

Swiss National Science Foundation, Universität Zürich and Swiss National Supercomputing Centre for financial support and computational time.

I would like to acknowledge the core electrons, without whom this work would not have been possible to realize.

Deep thanks goes to my colleagues, present and past, friends and relatives. I was luckily surrounded and blessed by far too many people to be individually mentioned here. Those persons contribute to keep coolness and provide invaluable support.

My special thanks to my nearest and dearest:
you're the color, you're the movement and the spin.

Curriculum vitæ

Samuele GIANI

Magdalenenstrasse 72, 8050 Zurich, Switzerland
+41(0)79/381.15.76
samuele.giani@gmail.com

Personal information

Citizenship	Swiss, Italian
Date and place of birth	16 Mai 1978 in Faido (TI), Switzerland
Civil status	Relationship, 2 children (born 2009 and 2011)

Education

05/2005 – 12/2013 **Universität Zürich**
Ph.D. in Physical Chemistry
– *Thesis: Calculations of spectroscopical properties of extended systems*
Advisor Prof. Dr. J. Hutter

10/1998 – 01/2005 **École Polytechnique Fédérale de Lausanne**
M.S. Molecular and Biological Chemistry
– *Diploma: Simulations of the enzymatic mechanism of hCAII*
Advisor Prof. Dr. U. Röthlisberger

09/1993 – 06/1998 **Scuola Cantonale di Commercio Bellinzona**
High School Maturità

Employment

10/2010 – Present	**Mettler Toledo Analytical AG** Application Chemist in Thermal Analysis Market Support Group for Material Characterization
05/2005 – 08/2010	**Universität Zürich** Teaching assistant: Physical Chemistry practical exercises, Informatics classes
06/2003 – 10/2003	**École Polytechnique Fédérale de Lausanne** Stage: Laboratoire d'Énergétique Industrielle
10/2001 – 06/2004	**École Polytechnique Fédérale de Lausanne** Teaching assistant: General Chemistry exercises

Professional qualifications

Languages	Italian (mother tongue), English, French and German
Informatics	Operating systems: Unix/Linux, MS Windows (MS Office), OS/X Programming languages: basis of `Fortran`, `C`, shell scripting, R, Perl, Python, LaTeX, vectorial Graphics, HTML
Hobbies and Sports	Playing bass guitar, chess, squash